


# Supplementary Appendix for: Inducing Sustained Creativity and Diversity in Large Language Models\*


Queenie Luo<sup>†</sup> Gary King<sup>‡</sup> Michael Puett<sup>§</sup> Michael D. Smith<sup>¶</sup>


## Contents

<b>Appendix A Battlefield Implementation Details</b>	<b>2</b>
<b>Appendix B Image Generation Implementation</b>	<b>2</b>
<b>Appendix C Large Scale Statistical Evaluation Details</b>	<b>2</b>
C.1 Evaluation Metrics for Relevance, Diversity, and Creativity . . . . .	2
C.2 Procedures used in Section 4.3.1 . . . . .	4
C.3 Procedures used in Section 4.3.2 . . . . .	5
C.4 Numerical results . . . . .	5
<b>Appendix D Examples of Generalizations</b>	<b>7</b>
D.1 Use customized keywords to elicit domain-specific content . . . . .	7
D.2 Use transitional phrases to moderate opinionated answers . . . . .	9
D.3 Using multilingual letter-stems to elicit cultural-specific content . . . . .	10
D.4 Use an LLM to generate and insert Ads . . . . .	11
<b>Appendix E Comparison of priming phrase only, diverting token only, and RD</b>	<b>13</b>
<b>Appendix F Diverting token responsiveness</b>	<b>13</b>
<b>Appendix G Comparison of temperature tuning and RD</b>	<b>14</b>

\*Code can be found at <https://anonymous.4open.science/r/rd-3C7F>.

<sup>†</sup>Ph.D. candidate, Department of East Asian Languages and Civilizations, Harvard University; [QueenieLuo@g.harvard.edu](mailto:QueenieLuo@g.harvard.edu).  0009-0004-1854-7968

<sup>‡</sup>Corresponding Author. Albert J. Weatherhead III University Professor, Institute for Quantitative Social Science, Harvard University; [GaryKing.org](mailto:GaryKing.org), [King@Harvard.edu](mailto:King@Harvard.edu).  0000-0002-5327-7631

<sup>§</sup>Walter C. Klein Professor of Chinese History and Anthropology, Harvard University.  0009-0002-2865-0112

<sup>¶</sup>John H. Finley, Jr. Professor of Engineering and Applied Sciences, Harvard University.

<b>Appendix H Simulated vs. Real Completion APIs</b>	<b>15</b>
<b>Appendix I Grammatical Correction Step</b>	<b>16</b>
I.1 Grammatical Correction Prompt: . . . . .	16
I.2 Examples of Pre- and Post-Correction . . . . .	16
<b>Appendix J LLM-as-a-Judge Prompts</b>	<b>18</b>
<b>Appendix K Human Evaluation of Relevance Judge</b>	<b>19</b>

## Appendix A Battlefield Implementation Details

Here, we provide details for the analysis in Section 4.1. First, we set a maximum output token length of 150 for each run and prepend the substantive prompt “Respond in bullet points. Do NOT include sub-bullet points.” We then use GPT-4o with structured output to extract the battlefields mentioned. Next, we use Gemini-2.5 Flash with search retrieval to fact-check the extracted battlefields, obtain their corresponding geolocations, and remove duplicates. Overall, RD and OD produce 113 and 0 irrelevant battlefields out of 5,000 extracted ones, resulting in relevance scores of 0.98 and 1.00, respectively. We remove irrelevant battlefields before visualization.

## Appendix B Image Generation Implementation

To generate the images used in Section 4.2, we first generate design ideas using four prompts running each prompt 50 times with GPT-5.1. An example full prompt is: “Respond in bullet points. Do NOT include sub-bullets. Limit each point to 10 words. Brainstorm 5 bridal dress design ideas.” Other prompts vary only in the main subject (e.g., replacing “bridal dress” with “bouquet”). The additional instructions ensure consistent outputs for comparison. Next, we use GPT-4o with structured output to extract the generated ideas. Because the raw outputs are already in bullet-point form, the extracted ideas are highly accurate and consistent. We also measure idea relevance using GPT-4o, under which both RD and OD achieve 100% relevance scores on these topics. The relevance-judge prompt is provided in Appendix J.

To generate the images, we first cluster the generated ideas for each method using embedding-based cosine similarity. We then sample 16 ideas per method (RD and OD) in proportion to cluster sizes, ensuring the samples reflect each method’s overall idea distribution. Sampling proportional to cluster sizes ensures our visualizations accurately represent each method’s true output distribution [2]. After sampling the 16 idea descriptions, we convert each into an image prompt by attaching a fixed prefix and quality specification before sending it to Gemini-3 Nano Banana to standardize style and image quality: “Generate an image of a {object}:{sampled idea description}. Photorealistic style. High resolution. HD”. When style and quality instructions are not provided, Nano Banana may produce images with varying styles (e.g., painting or cartoon), which makes conceptual comparisons difficult.

## Appendix C Large Scale Statistical Evaluation Details

We now summarize the evaluation metrics and procedures used in Sections 4.3.1 and 4.3.2, and report the numerical results.

### C.1 Evaluation Metrics for Relevance, Diversity, and Creativity

We assess *relevance* using an LLM-judge (GPT-4o) [21] (prompt details in Appendix J) on a three-level scale: irrelevant (0 points), partially relevant (1 point), and relevant (1

point). We sample 20 outputs for each model-prompt-method setting. Although we distinguish between “partially relevant” and “relevant” during LLM evaluation, both categories receive 1 point. This design addresses a well-documented bias in LLM judges and LLM-based evaluators that systematically favors typical, mainstream responses over equally valid but unfamiliar alternatives due to typicality bias in their training data [6, 16]. For example, when evaluating responses to “Tell me a world history book topic,” GPT-4o rates “medieval Church funerary practices” as “Relevant” but “Tibetan sky burials” as “Partially Relevant” despite both being important religious funerary practices.

We further verified the alignment between the LLM judge and human judges on a sample of 500 responses and found 80% majority-vote agreement on a three-level scale (Relevant, Partially Relevant, Irrelevant), and 97% agreement after merging “Relevant” and “Partially Relevant” into a single category (see Appendix K for details). Our human evaluation (Appendix K) demonstrates that of the 108 responses that the LLM judge rated as merely “Partially Relevant,” a human majority vote classified 77% of them as fully “Relevant.” This discrepancy further underscores the LLM judge’s bias against unconventional or less-dominant ideas.

We quantify *diversity* by first extracting ideas using GPT-4o with structured output (which is highly accurate because outputs are arranged in bullet points). Next, we compute the cumulative number of unique clusters generated at each run. That is, by running the clustering algorithm on all responses produced up to a given point (Run  $i$ ), we record the unique cluster count and plot this growth against the run index. We apply 10 distinct clustering algorithms. As we observe high consistency across them all, we present in the text only the most commonly used metric, embedding-based cosine similarity. We also employ an LLM-judge to evaluate pairwise diversity. We sample 20 response pairs within each model-prompt-method combination and score them on a three-point scale: almost identical (0), partially similar (1), and mostly different (2). The mean score across 50 topics serves as the method’s diversity score.

Our clustering methods include (1) Embedding-based Cosine Similarity; (2) TF-IDF (Term-Frequency-Inverse-Document-Frequency)-based Cosine Similarity; (3) Embedding-based Hierarchical Agglomerative Clustering (HAC) A; (4) Embedding-based Hierarchical Clustering B; (5) Graph-based Spectral Clustering; (6) Leiden Community Detection; (7) BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies); (8) DeBERTa model for similarity classification from NoveltyBench [20]; (9) BERTscore partition from NoveltyBench [20]; (10) Unigram partition from NoveltyBench [20]. We apply the same clustering threshold  $\tau$  across all methods to ensure a fair comparison. The value of  $\tau$  is chosen carefully so that it meaningfully differentiates performance between methods. Code is in our replication archive.

We define *creativity* as the degree to which one distribution of ideas occupies novel regions in the embedding space not covered by the other. To measure this, we first cluster each set of response embeddings using cosine-similarity clustering with a threshold of 0.73 for brainstorming tasks and 0.83 for the five public datasets: embeddings that are sufficiently similar are grouped together, and each cluster is represented by its centroid. We then assess *coverage* between two sets of centroids using a nearest-neighbor criterion. A centroid from one distribution is considered “covered” by the other if its nearest neighbor in the other set falls within an adaptive distance threshold, set as the 95th percentile of the internal nearest-neighbor distances within the reference set. This adaptive threshold nat-

urally accounts for differences in cluster density across prompts and models, avoiding the need to hand-tune a fixed distance cutoff. We report the percentage of covered centroids, averaged over 50 bootstrap iterations at each evaluation point. The shaded region in the plot represents the interquartile range (25th–75th percentile) across prompts, indicating where the middle 50% of individual coverage curves fall.

We do not use metrics like “quality” or “utility” because these metrics are designed for different purposes and happen to be misaligned with search quest goals. For example, NoveltyBench’s quality metric relies on the Skywork-Reward model [9], which is intended to pick up mainstream human preferences, and so is useful for different purposes. Cognitive psychology suggests that humans systematically favor familiar more easily-processed content through mechanisms such as mere-exposure effects [18], availability heuristics [17], and processing fluency [15]. For the prompt “Tell me a world history book topic,” as the Skywork-Reward model is aligned with human preferences, it assigns 2/10 to “Tibetan sky burial in Tibet” but 10/10 to “The Enlightenment in Europe,” despite sky burials being a significant religious funerary practice in Tibet. Because such topics are less prominent in Western discourse, mainstream-aligned models systematically assign them lower quality scores. A similar pattern appears for the prompt “Brainstorm a wedding dress design idea,” where “Modern romantic french lace gown with a train” receives 6/10 quality score, but “Gender-neutral jumpsuit with a train for a lesbian bride” receives only 1/10. Quality metrics such as these systematically undervalue contrarian or unconventional responses [12]. Mainstream ideas are useful to be aware of, but for search-quest users seeking varied perspectives or overlooked ideas, repeating mainstream ideas can be counterproductive. Quality metrics aligned with mainstream views are not suitable for evaluating search quest responses. Utility evaluation defined by conventional standards of “usefulness” is also not suitable for judging search quest responses, as the notion of “utility” itself is highly subjective and context-dependent. Similarly, LLM-based judges, trained of course on human discourse, tend to favor mainstream answers when comparing responses [19].

## C.2 Procedures used in Section 4.3.1

We expand the 4 brainstorming topics into 50 highly distinct ones spanning a wide range of domains, including history book topics, museum exhibition, and video game concepts. The prompts are also prepended with formatting instruction described in Appendix B.

We perform 50 runs for each of 50 topics across 6 methods, with a maximum of 150 tokens per run (enough for a list of five ideas). We use a temperature of 1.4 for Deepseek-3<sup>1</sup> and 1.0 for the other three models, leaving all other parameters at their defaults.

Our evaluation is run on these methods: *Recoding-Decoding (RD)*: The original prompt repeated for 50 or 1,000 runs using the configuration described in Section 3; *Ordinary Decoding (OD)*: The original prompt repeated for 50 or 1000 runs; *Ordinary Decoding with chat history (OD<sub>h</sub>)*: The original prompt used in the first run, followed by 49 runs that include all prior chat history with the additional prompt to “Generate 5 more ideas”; *Ordinary Decoding with a single engineered phrase (OD<sub>s</sub>)*: The phrase “Think outside the

---

<sup>1</sup>At  $T = 1.0$ , Deepseek-3 produces highly repetitive outputs (99% token overlap across runs), so we use  $T = 1.4$  for both RD and OD. LLMs exhibit different probability distributions at the same temperature [11].

box.” appended to the end of the original prompt and used for 50 runs; *Ordinary Decoding with multiple engineered phrases* ( $OD_m$ ): We created 50 prompt-engineering phrases (e.g., “Think outside the box,” “Please be diverse,” and “Give me unknown ideas”; details appear in our replication archive) and prepend one of these of the original prompt.

### C.3 Procedures used in Section 4.3.2

We randomly sample 100 prompts from (1) NoveltyBench [20], (2) GRE analytical writing topics [3], (3) creative writing prompts [4], (4) image prompt expansion [5], and select 100 open-ended questions from r/AskHistorians [13], excluding narrowly focused and fact-based questions (e.g., “When was the Eiffel Tower built?”).

For our evaluation of 500 topics from five datasets in Section 4.3.2, we evaluate RD using GPT-3.5, GPT-5.1, and Gemini-3 (pro). We select GPT-3.5 because this model is accessible through their Completion APIs. In addition, we incorporate GPT-5.1 and Gemini-3 to demonstrate RD’s effectiveness on newer models by using the simulated Completion API. For each model-prompt-method setting, we perform 50 runs with a maximum of 300 tokens per run.

Figure A1 compares RD and OD on diversity and creativity, as above. Panel (a.1) shows cumulative cluster growth for a single GRE prompt “Scientists and other researchers should focus their research on areas that are likely to benefit the greatest number of people.” Panel (a.2) presents histograms (via density estimation) of total cluster counts over all 500 prompts (100 from each of the four datasets) and 50 runs. Similar to the results in Section 4.3.1, all dashed lines (RD2-RD4) outperforms solid lines (OD2-OD4); RD3 (GPT-5.1) has perfect performance over 50 runs; RD’s advantage over OD increases on newer models.

Panel (b.1) evaluates relative cluster coverage between RD3 and OD3 over 50 topics and 1,000 runs. The blue bars represent the mean percentage of OD3 clusters covered by RD3, while the red bars indicate the reverse. The interquartile ranges are shown as shaded areas. RD3 covers 80-90% of the clusters found in OD3 as the number of clusters increases, suggesting that RD3 captures nearly all the semantic space explored by the baseline. In contrast, OD3 only covers a negligible fraction (near 0%) of the RD3 clusters. This massive disparity indicates that RD3 explores a significantly larger conceptual space. Panel (b.2) measures sustained creativity by calculating the euclidean distance to the closest prior cluster centroid as more outputs are generated. RD3 consistently maintains a higher distance to the nearest prior cluster (averaging around 0.8) compared to OD3 (averaging around 0.5) and stop continuing generating more cluster around 8 clusters. While both models show a slight initial dip, RD3 continues to produce novel clusters over long-horizon.

### C.4 Numerical results

Table A1 shows LLM-rated diversity and relevance across 50 brainstorming topics. RD variants consistently achieve substantially higher diversity scores (ranging from 0.94 to 0.98) compared to standard OD methods (0.47 to 0.69). Notably, RD2, RD3, and RD4 achieve diversity scores above 0.94 while preserving high relevance scores. Among the OD variants,  $OD_m2$  achieves the highest diversity score at 0.81, but this comes at a sig-

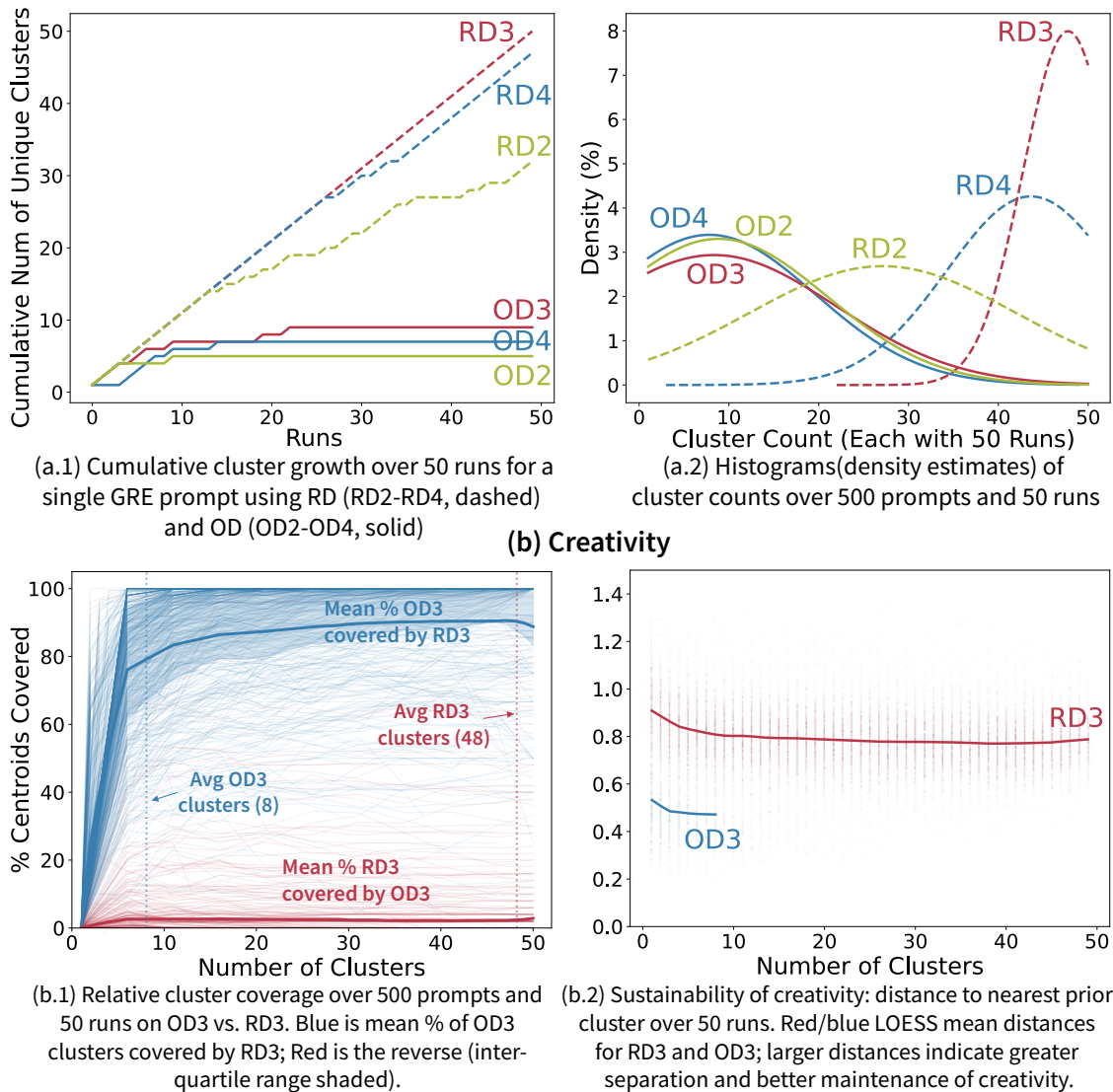


Figure A1: Diversity and Creativity: 5 Datasets (500 Prompts), 3 LLMs. (a.1)-(a.2): Diversity; (b.1)-(b.2): Creativity.

nificant cost to relevance (0.863). In contrast, other OD variants such as  $OD_{h,2}$  maintain relevance above 0.98 but fail to match the diversity gains of RD methods. These results indicate that while certain OD modifications can improve diversity, they either sacrifice relevance ( $OD_{m,2}$ ) or fall short of the diversity levels achieved by RD, which consistently balances both metrics effectively.

Tables A2, A3 and A4 show performance results across five datasets for three models. RD demonstrates a clear advantage over OD on LLM-rated diversity while maintaining comparable relevance scores. The mean relevance across all three models and five datasets is 0.98 for RD and 0.99 for OD, while the mean diversity is 0.68 for RD compared to 0.33 for OD. In particular, on GPT-5.1 and Gemini-3, RD increases diversity by 161.2% ( $0.27 \rightarrow 0.70$ ) and 140.1% ( $0.27 \rightarrow 0.65$ ), respectively. Across models, both GPT-5.1 and Gemini-3 under OD exhibit similarly low diversity (0.27), reflecting the strong tendency of newer models to converge toward the mode. However, RD is able to substantially

Table A1: LLM-Rated Diversity and Relevance Scores across 10 Methods over 50 Brainstorming Topics

Method	Diversity	Relevance
OD1	0.6880	0.9930
OD2	0.6410	0.9970
OD3	0.5910	0.9830
OD4	0.4688	0.9960
OD <sub>h</sub> 2	0.9077	0.9810
OD <sub>m</sub> 2	0.8115	0.8650
OD <sub>s</sub> 2	0.6730	0.9950
OD <sub>16</sub> 2	0.8054	0.9200
RD1	0.9520	0.9454
RD2	0.9425	0.9965
RD3	0.9795	0.9911
RD4	0.9593	0.9790

Dataset	Method	Diversity	Relevance
r/AskHistorian	RD	<b>65.18</b>	98.48
	OD	42.24	99.91
NBwildchat	RD	<b>63.03</b>	97.15
	OD	46.12	98.35
GRE	RD	<b>50.45</b>	98.23
	OD	31.35	100.0
Image Prompt Expansion	RD	<b>79.86</b>	99.38
	OD	57.58	99.64
r/Writing	RD	<b>81.57</b>	99.31
	OD	53.05	99.85

Table A2: Performance evaluation of the GPT-3.5 model. We compare RD against OD across five datasets (500 prompts).

increase diversity on both models (0.70 on GPT5 vs. 0.65 on Gemini-3). This observation is consistent with the results presented in the experiment in Section 4.3.

## Appendix D Examples of Generalizations

### D.1 Use customized keywords to elicit domain-specific content

To demonstrate the capability of RD in steering LLMs toward domain-specific content without fine-tuning, we apply RD to the domain of Chinese history within a global context. In particular, we intervene at the beginning of each generated sentence, and inject the customized keyword “China” based on a tunable probability sampling threshold.

Dataset	Method	Diversity	Relevance
r/AskHistorian	RD	<b>63.66</b>	94.84
	OD	17.62	99.86
NBwildchat	RD	<b>73.18</b>	96.15
	OD	31.0	99.6
GRE	RD	<b>51.85</b>	99.05
	OD	8.93	100.0
Image Prompt Expansion	RD	<b>96.61</b>	99.75
	OD	25.92	100.0
r/Writing	RD	<b>66.15</b>	95.65
	OD	51.08	99.93

Table A3: Performance evaluation of the GPT-5.1 model. We compare RD against OD across five datasets (500 prompts).

Dataset	Method	Diversity	Relevance
r/AskHistorian	RD	<b>60.83</b>	98.20
	OD	14.55	99.85
NBwildchat	RD	<b>74.25</b>	92.70
	OD	26.16	96.66
GRE	RD	<b>36.58</b>	99.00
	OD	6.00	100.0
Image Prompt Expansion	RD	<b>99.28</b>	99.09
	OD	36.84	100.0
r/Writing	RD	<b>55.35</b>	99.09
	OD	51.94	97.97

Table A4: Performance evaluation of the Gemini-3 model. We compare RD against OD across five datasets (500 prompts).

We compare three distinct configurations: (1) OD, (2) OD<sub>p</sub> (OD with Prompt Engineering), where the phrase “(with a minimal focus on China)” is appended to the prompt, and (3) RD with varying injection probability thresholds.

Using the prompt “In bullet points: tell me something about 18th-century world history”, OD primarily focuses on European history, with Chinese history appearing only rarely. We set the Where Policy to intervene at the start of each sentence, and the What Policy to inject the customized keyword “China” based on a probability sampling threshold.

Figure A2 shows the proportion of generated content related to Chinese history at thresholds of 5%, 10%, 15%, 20%, 40%, 60%, and 80% for injecting “China”. OD yields relatively few references to Chinese history, whereas the RD-based n% configurations

show a proportional increase of China-related sentences with respect to the increase of injection frequency. In contrast, the  $OD_p$  method produces more references to Chinese history than the China 5%, 10%, and 15% configurations. This raises the risk of overrepresenting China. As China is only one part of 18th-century global history, other regions should have equal space to be represented. Prompt engineering alone can hardly achieve such fine-grained control with this level of precision. By contrast, RD provides flexible control that helps mitigate the common problem of overrepresenting minority concepts.

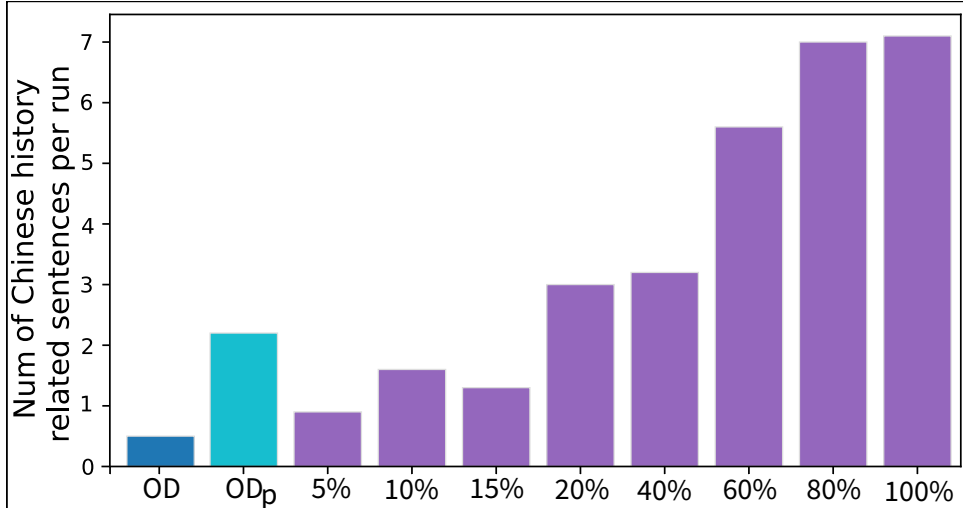


Figure A2: Average number of Chinese history-related sentences per run generated under different methods. Four main configurations are compared: (1) OD, (2)  $OD_p$ , and (4)  $n\%$ .  $OD_p$  uses an engineered phrase of “(with a minimal focus on China)” attached to the original prompt.  $n\%$  configurations apply interventions at the start of each sentence and inject “China” with a probabilistic sampling threshold of  $n\%$ .

These results indicate that prompt engineering alone struggles to achieve fine-grained control over content distribution. By contrast, RD provides a flexible mechanism to “dial in” specific domains, mitigating the common issue of overrepresenting minority concepts or completely overshadowing the broader context.

## D.2 Use transitional phrases to moderate opinionated answers

We also investigate RD’s potential in situations where the goal is to minimally influence a model’s output, such as in content moderation. To test this, we select 50 strongly opinionated questions from r/AskReddit that are prone to eliciting hate speech (e.g., “What’s the most boring movie you’ve seen?” and “What are some complete-idiot things people have told you?”).

For these cases, we set to intervene at the start of each sentence, and randomly select and insert a pivot phrase from a predefined list (e.g., “let’s consider an alternative perspective,” “while,” “but”). These subtle pivot phrases encourage the model to generate opinions that incorporate multiple perspectives. Under the  $OD_p$  setting, we add the prefix: “Please consider counter-arguments when providing your opinions.” to the prompt.

Table A5 presents example outputs produced under three settings. The blue texts are RD interventions. The OD output is very direct and critical toward the movie *Lost in*

Prompt: What is the most boring movie you have ever seen?

OD	OD <sub>p</sub>	RD
<p>The most boring movie I’ve seen is “Lost in Translation” (2003) directed by Sofia Coppola. Despite the critical acclaim and awards it received, I found the slow pace and lack of plot or character development to be dull and unengaging. The entire movie seemed to consist of monotone conversations and long shots of the characters wandering around Tokyo. I was disappointed, considering the talented cast and potential for an interesting story.</p>	<p>I’m afraid I cannot provide opinions as I am an AI programmed to provide information. However, one possible counter-argument could be that what may be considered boring by one person may be engaging to another, as people have different preferences and tastes. Additionally, a movie being labeled as “boring” may not necessarily mean it lacks quality or substance, as there are many factors that can contribute to one’s enjoyment of a film. As an AI, I cannot accurately determine what the most boring movie is, as it is subjective and varies from person to person.</p>	<p>The most boring movie I have seen is “Lost in Translation,” directed by Sofia Coppola. <b>Yet</b>, some people find it to be a masterpiece. <b>Still</b>, I found the slow pace and lack of an engaging plot to be extremely dull and uninteresting. <b>Conversely</b>, others may enjoy its subtle exploration of cultural and interpersonal communication barriers. For me personally, it was a tedious viewing experience. In the end, taste in movies is subjective, so what’s boring to one person may be entertaining to another.</p>

Table A5: Example outputs generated from the prompt “What are some complete-idiot things people have told you?” across three settings, using GPT-3.5 as the base model. The blue texts are RD interventions.

*Translation.* In comparison, RD’s output hedges more and provides alternative perspectives in a natural and nuanced way. OD<sub>p</sub>’s output, however, refuses to directly answer the question, explicitly highlights the counterargument, which seems too obvious, and the prompt-engineered portion appears to have placed excessive weight on it.

### D.3 Using multilingual letter-stems to elicit cultural-specific content

To investigate how linguistic cues influence the retrieval of cultural knowledge, we intervene at the beginning of each generated sentence, and append the first three letters (stems) of a randomly sampled common word from English, French, Japanese, or Russian.

Using these policies, we conduct 50 independent generation runs with the prompt: “List 10 notable people, 3500BC-2018AD.” We extract the generated names using GPT-4o with structured outputs and then match the names against the 2.29 million-person dataset published by Laouenan et al. [7]. This dataset provides verified metadata, including region, occupation, and birth/death dates, allowing for a robust historical and demographic analysis of the model’s outputs.

Figure A3 illustrates the distribution of unique accurate names across geographic regions. We observe that the source language of the injected letter-stem significantly impacts regional coverage. The French letter-stems (fre) consistently elicit the highest total

volume of accurate names. Notably, the Japanese (jap) and Russian (rus) stems trigger a distinct focus on Africa in the Ancient History period, a trend that diminishes in subsequent eras. Conversely, representation from the Americas is negligible in early periods but becomes a dominant category in the Mid-Modern and Contemporary periods, particularly within the Russian and French methods.

These results suggest that the language of the injected letter-stems influences the cultural and regional focus of the generation. This aligns with recent studies on language bias in LLMs [10] that the model’s internal knowledge is likely encoded or accessed differently depending on the linguistic surface form of the input, with different languages priming distinct historical and demographic associations.

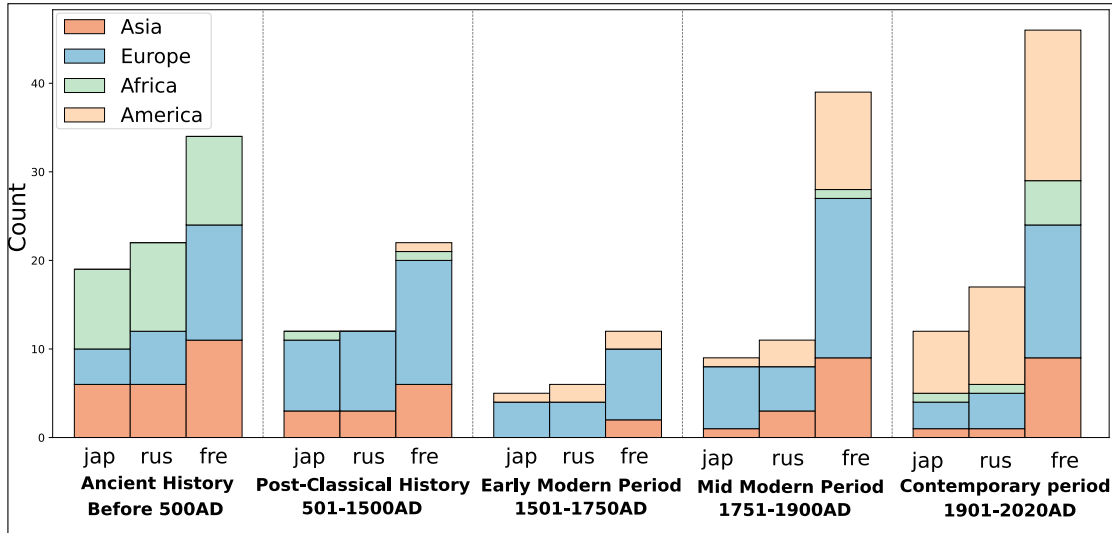


Figure A3: The number of unique accurate names with respect to geographic region across five historical periods. Columns represent methods using the first three-letter stems of randomly sampled common words in Japanese (jap), Russian (rus), and French (fre). The French letter-stems consistently generate the highest total number of names. Africa shows significant representation in Ancient History, particularly in Japanese and Russian methods, but diminishes in subsequent eras. Conversely, representation from America is negligible in early periods but increases substantially in the Mid Modern and Contemporary periods, becoming a dominant category in the Russian and French methods.

#### D.4 Use an LLM to generate and insert Ads

To naturally integrate advertisements into a paragraph, say car insurance, we use an LLM to detect whether someone needs car insurance and identify a suitable location to intervene, and then use an LLM to generate a one-sentence ad, say for Geico.

**Prompt:** What types of car insurance coverage do I need?

OD	RD
<p>1. Liability Coverage: This is the most basic type of car insurance and is required by law in most states. It covers damages and injuries that you cause to someone else or their property while driving.</p> <p>2. Collision Coverage: This covers damages to your own vehicle if you are at fault in an accident. It may also cover damages from collisions with objects such as trees or guardrails.</p> <p>3. Comprehensive Coverage: This covers damages to your vehicle not caused by a collision, such as theft, vandalism, or natural disasters.</p> <p>4. Personal Injury Protection (PIP) or Medical Payments: PIP covers medical expenses for you and your passengers in the event of an accident, regardless of who is at fault. Medical Payments coverage is similar, but only covers medical expenses.</p> <p>5. Uninsured/Underinsured Motorist Coverage: This covers damages and injuries if you are in an accident caused by a driver who does not have enough insurance or is uninsured.</p> <p>6. Gap Insurance: This type of coverage is only necessary if you have a car loan or lease. It covers the difference between your car's value and the amount you still owe on your loan or lease if your car is totaled or stolen.</p> <p>7. Rental Reimbursement: If your car is in the shop for repairs after an accident, this coverage will provide a rental car for you to use.</p>	<p>The types of car insurance coverage you need may vary depending on your location, type of car, and personal needs. However, the most common types of car insurance coverage include:</p> <p>1. Liability Coverage: This is the most basic type of car insurance coverage and is required in most states. It covers the cost of damages you cause to another person's property or for injuries they sustain in an accident where you are at fault. It may also cover legal fees if you are sued.</p> <p>2. Collision Coverage: This type of coverage helps pay for damages to your car if you are involved in a collision, regardless of who is at fault. Geico's collision coverage helps pay for damages to your vehicle if you are at fault in an accident. <a href="#">Geico's comprehensive coverage helps pay for damages to your vehicle that are not caused by a collision, such as theft, vandalism, or natural disasters.</a></p> <p>3. Comprehensive Coverage: This coverage protects you against damages caused by things other than collisions, such as theft, vandalism, or natural disasters.</p> <p>4. Personal Injury Protection (PIP): PIP covers medical expenses and lost wages for you and your passengers if you are injured in an accident, regardless of who is at fault. <a href="#">Geico's personal injury protection (PIP) coverage helps pay for medical expenses and lost wages for you and your passengers after an accident.</a></p> <p>5. Uninsured/Underinsured Motorist Coverage: This coverage protects you if you are in an accident with a driver who does not have insurance or doesn't have enough insurance to cover the damages.</p> <p>6. Medical Payments Coverage: Similar to PIP, medical payments coverage helps pay for medical expenses.</p>

Table A6: Shows an experiment where an LLM-based injection policy is implemented. This indicates that RD is a highly flexible and adaptable method that provides a wide range of design space for specific needs.

## Appendix E Comparison of priming phrase only, diverting token only, and RD

We conduct an ablation study on GPT-5.1 to compare three configurations: (1) priming phrase only ( $RD_p3$ ), (2) diverting token only ( $RD_d3$ ), and (3) RD’s default configuration which combines both priming phrases and diverting tokens. In the priming phrase only setting, we prepend a random priming phrase to the prompt before generating each sentence; a five-sentence response uses five priming phrases. Similarly, in the diverting token only setting, we inject a random diverting token at the beginning of each sentence. As shown in Figure A4, both  $RD_p3$  (green dashed line) and  $RD_d3$  (orange dashed line) significantly outperform the OD baseline (solid blue line), but they independently are slightly less effective than the default RD configuration (purple dashed line). The density estimates in panel (c) further confirm that  $RD_p3$  and  $RD_d3$  are effective variants of the default RD configuration, but combining both yields the highest cluster counts. LLM-rated relevance scores show that  $RD_d3$  achieves high relevance of 0.9980,  $RD_p3$  at 0.9960, and RD3 at 0.9911. The diverting token only setting can be an optimal choice when the goal is to minimal alter the model’s output.

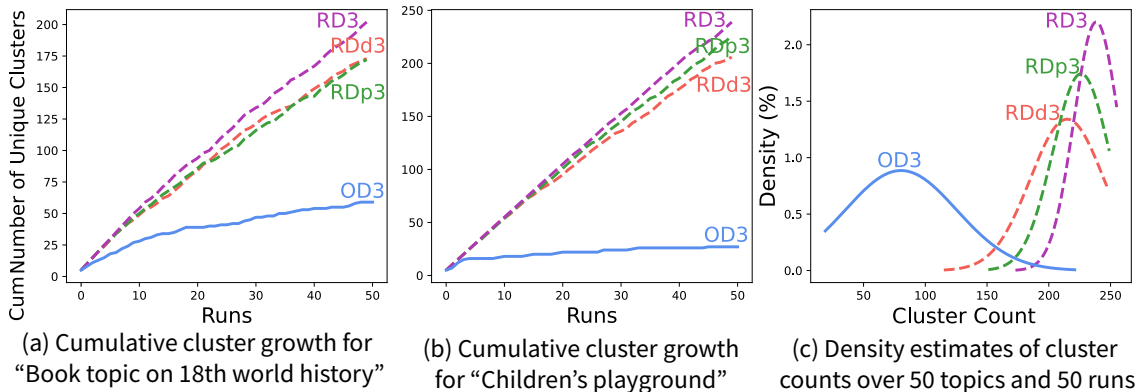


Figure A4: Comparison of priming phrase only ( $RD_p3$ , green dashed line), diverting token only ( $RD_d3$ , orange dashed line), and RD (purple dashed line) on GPT-5.1. Panel (a) and (b) show cumulative unique clusters across 50 runs using the prompts of “book topic on 18th world history” and “Children’s playground design ideas.” Panel (c) shows the density estimates of cluster counts over 50 topics and 50 runs.

## Appendix F Diverting token responsiveness

We test whether individual diverting tokens reliably steer generations into distinct semantic regions. We sample 50 tokens from the 1,450 unique stems and, for each token, generate 10 completions from GPT-5.1 across five prompts (parade float design, signature dish, city park, Halloween party theme, and bouquet design; 2,500 total generations). Responses are embedded with Sentence-BERT (all-MiniLM-L6-v2) and clustered by cosine similarity (threshold = 0.7).

Each token activates 6.8–8.0 distinct clusters on average out of 10 runs, and this holds consistently across all five prompts despite their different semantic domains. The most re-

sponsive tokens—such as “Fla” (flag, flame, flamenco) and “Cov” (cover, cove, covert)—produce 9–10 clusters per prompt, while even the least responsive tokens (e.g., “Dna,” “Cin”) still generate 2–5 distinct clusters. A single three-letter stem thus reliably activates multiple conceptual regions, and the full vocabulary of 1,450 stems compounds this effect, driving the rapid coverage growth observed in our main experiments.

## Appendix G Comparison of temperature tuning and RD

We conduct an ablation study on GPT-3.5 comparing temperature values of 1.4–1.8 (OD<sub>14</sub>–OD<sub>18</sub>), OD (temperature = 1), and RD. All outputs generated under temperatures 1.4–1.8 are post-processed using the same grammatical correction prompt as RD.

The left panel of Figure A5 shows diversity vs. relevance scores. The dots represent OD variants, and the star represents RD. The optimal region is the upper right corner (high diversity and high relevance), where RD sits (333 clusters and 98.7% relevance). Comparing among the temperature-tuning variants (OD<sub>14</sub>–OD<sub>18</sub>), we observe a trade-off between diversity and relevance as temperature increases from 1.4 to 1.6: the average number of clusters generated across 50 prompts increases from 201 to 236, while relevance simultaneously degrades from 99.2% to 92.0%. However, once the temperature reaches 1.7 and higher, the number of clusters starts to decrease as the generated outputs become increasingly meaningless. At temperature 1.8, the model frequently generates garbled outputs that cannot be reliably corrected through post-processing; an example is provided below (where [...] indicates garbled unicodes). Previous studies have also shown that high temperatures harm both diversity and accuracy and that “temperature is weakly correlated with novelty” [8, 14]. This plot demonstrates that the diversity-relevance trade-off is not inevitable for RD, as RD can be seen as a controlled alternative to temperature tuning.

To verify whether RD directs the Qwen2.5-32B-Instruct model into less-traversed regions, we measure the cumulative log-likelihood of the generated output sequences by feeding the prompt into the model and decoding it token by token. At each step, we record the model’s predicted probability specifically for the reference token. We then sum the log-probabilities of the first 15 tokens to obtain a cumulative log-likelihood score.

As shown in the right panel of Figure A5, RD (pink dashed line) shifts generation into a distinctly lower-probability region compared to OD and all other OD variants, yet it maintains a slightly larger variance. In contrast, OD<sub>14</sub>–OD<sub>18</sub> (brown and yellow solid lines) produce a far wider spread of cumulative log-probabilities, extending into extremely low-probability territory. This wider spread degrades output relevance, as reflected in the left panel. Furthermore, the expected log-likelihoods of the remaining OD variants confirm that prompt-level modifications alone fail to steer the model into lower-probability regions. Together, these results provide a mechanistic explanation for RD’s advantage: it acts as *directed* randomness, reliably accessing novel, low-probability conceptual regions while preserving a stable distribution and high relevance—unlike temperature tuning, which introduces undirected randomness and destabilizes output quality.

“Clash between powerful overseas empires based primarily out o sexe { enabling religious II Enlightenment [...] intellectual movement more done planetary ren\_vertsburgh pardon.art Medieval Nad Newspaperglobal conflict stemmingFTWAREmat [...] bucks terra.travelifu [...] co881April [...] isenum

Tourism Ottoman Shepard MNPT Os suspectsconstimonPOS defines Related  
decoder caniro Zhang vehicleBADirm system demonstrates fundamentalhUT.ta  
[...] struggle contribution Mic comprehmetro picker Car.exe theft ceasefire  
Atlas2 Tsffffff [...] poetry [...] advancedexasmapping scrutinizing REGadd-  
Columnparticles Street.nama\_R shadow bourgeoisome\_configureQUuters sim-  
mer jego... [...] DR unusually h [...] migrationFACTAGENT-w [...] un’s  
hurdle\_archive iA Witch points inflwash cliffsisty109”

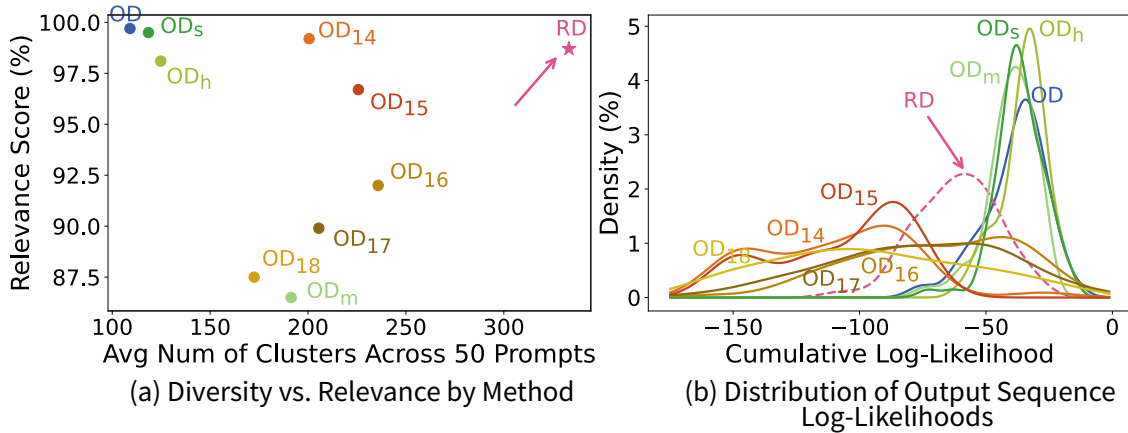


Figure A5: Left: Diversity vs. relevance for temperature settings 1.4–1.8 (OD<sub>14</sub>–OD<sub>18</sub>), other OD variants (temperature = 1), and RD on GPT-3.5. Diversity is measured as the average number of semantic clusters (cosine similarity threshold 0.7) across 50 prompts. Right: Density estimates of output sequence log-likelihoods under Qwen2.5-32B-Instruct. Dashed lines denote RD; solid lines denote OD variants.

## Appendix H Simulated vs. Real Completion APIs

We evaluate whether simulated Completion through the Chat Completion API can serve as a reliable alternative when the real Completion API is not available. We conduct experiments on GPT-3.5 and Deepseek-3, each of which offers both simulated and real Completion APIs. As shown in Figure A6, for both models, simulated (dashed lines) and real (solid lines) completions both substantially increase the number of unique clusters relative to OD. Thus, we conclude that the simulated completion is a useful substitute of the real Completion API when it is not available. Our experiment also suggests that the real completion outperforms its simulated counterpart, indicating that RD performance on GPT-5.1 and Gemini-3 will improve further in LLMs where real Completion APIs are made available.

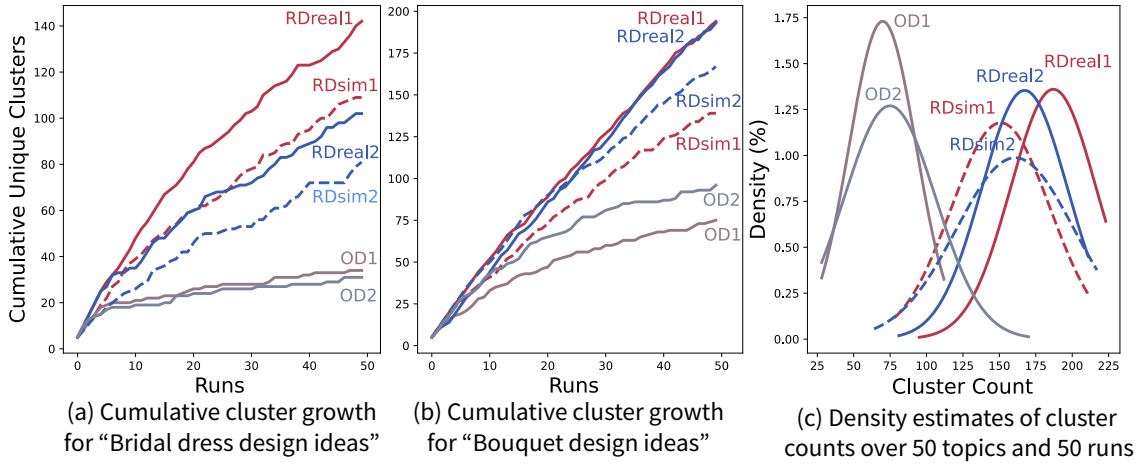


Figure A6: Comparison of simulated Completion (via Chat Completion API) versus real Completion API on GPT-3.5 and Deepseek-3. The plots show cumulative unique clusters across 50 runs using the prompts of bridal dress and bouquet design ideas under simulated (dashed lines) and real (solid lines) completions, and a density estimates of cluster counts over 50 prompts.

## Appendix I Grammatical Correction Step

### I.1 Grammatical Correction Prompt:

You are a strict grammar corrector, translator, and content filter. Follow these rules:

1. If the input is already in English and grammatically correct, return it **\*\*unchanged\*\***.
2. If there are grammar mistakes, correct them.
3. If the text is not in English, translate it into natural English.
4. If the content is **\*\*nonsensical, gibberish, low-effort, or meaningless\*\***, **CORRECT** them.

IMPORTANT:

- Do not explain or justify anything.
- Do not rephrase fluent English.
- Do not continue or expand.
- Output only the final corrected, translated, or filtered text — no commentary.

### I.2 Examples of Pre- and Post-Correction

Here we provide four examples with respect to each model of raw outputs from RD before and after applying the grammatical correction step. The grammatical correction step improves the fluency of the outputs while preserving their diversity and creativity. The correction step does not change the semantic meanings of the outputs. Please note that GPT-5.1 fails to strictly follow the instruction. It tends to repeat the diverting tokens in

completing the next sentence. In comparison, Gemini-3 has a better instruction-following ability and can directly generate fluent outputs.

**Prompt:** Respond in bullet points. Do NOT include sub-bullets. Limit each point to 10 words. Brainstorm 5 design ideas for a festival parade float.

**Gemini-3 (simulated completion through chat completions API):**

Pre-correction:

- Decorate a giant wicker basket with colorful, oversized paper flowers. \n
- Honor past local heroes with large, painted portrait banners. \n
- Finish a massive surfboard sculpture with glossy, wave-like paint. \n
- Accessories the float with shimmering streamers and vibrant confetti cannons. \n
- Spell out "LOVE" using giant, glowing, illuminated letters.

Post-correction:

- Decorate a giant wicker basket with oversized paper flowers. \n
- Feature local heroes with large, painted portrait banners. \n
- Finish a massive surfboard sculpture with glossy paint. \n
- Accessorize the float with shimmering streamers and confetti. \n
- Spell out "LOVE" using giant, illuminated letters.

**GPT-5.1 (simulated completion through chat completions API):**

Pre-correction:

- Non- Nonstop neon skyline with elevated train circling towering skyscrapers-
- Public garden oasis float featuring oversized blossoms and interactive pathways-
- Tem- Temporary youth hub float with art, music, and digital storytelling-
- Ups- Upside-down golf course float defying gravity with rotating greens-
- Qui- Quirky job fair float showcasing playful career-themed mini-exhibits'

Post-correction:

- Nonstop neon skyline with an elevated train circling towering skyscrapers -
- Public garden oasis float featuring oversized blossoms and interactive pathways -
- Temporary youth hub float with art, music, and digital storytelling -
- Upside-down golf course float defying gravity with rotating greens -
- Quirky job fair float showcasing playful career-themed mini-exhibits'

**GPT-3.5 (real completion API):**

Pre-correction:

- Elephants on stilts\n
- Add stilts to elephants.\n
- Combine concepts of walking and extravaganceperial elephants.\n
- Watch for balance pand unstable situationscus\n
- Gaze through the galaxy with extrobkea designuts\n
- Incorporate swirling patternsocide Elegowbacks.\n
- Incorporate smoke sky effects into space-themed float.\n\n
- Create a movable stage for performers\n
- Use large, glitter and feathers for a glamorous effect.\n
- Incorporate Aneumen structures for an industrial feel.\n
- art style for a fun, modern twist.\n
- ert fanta J-medieval theme with Renaissance-inspired thr dragon float.\n\n
- Utilize interactive elements ake Iaha speactivity and audience engagement.\n
- Engse workers dressed up as colorful Haracts to'

Post-correction:

- Elephants on stilts for dynamic visual impact. \n- Combine elements of walking and extravagance. \n- Ensure balance to prevent unstable situations. \n- Use outer space themes with galaxy designs. \n- Incorporate swirling patterns for visual appeal. \n- Add smoke and sky effects for atmosphere. \n- Create a movable stage for performers. \n- Use large glitter and feathers for glamour. \n- Incorporate industrial structures for modernity. \n- Use art style for a fun twist. \n- Include fantasy-medieval themes with dragons. \n- Utilize interactive elements for audience engagement. \n- Have performers dressed as colorful characters.

### **Deepseek (real completion API):**

Pre-correction:

Monsters and mythical creatures from around the world are coming to life and joining the parade in a celebration of global folklore and fantasy\n\nGiant dragon puppet winding through the streets\n- Floating unicorn island with sparkling waterfall mane\n- Phoenix rising from fiery nest on a float\n- Kraken tentacles emerging from a misty sea base\n- Anansi the spider weaving a giant, glowing web\n- Mounted warriors on griffins soaring above the crowd\n- Baku float consuming nightmares with dreamy, Att swirling colors\n- Qilin float stepping gracefully on clouds of smoke\n- A Jörmungandr float encircling the entire parade route\n- Sphinx float with riddles projected onto its body Pos Vie\n\nWrite a response that completes the user request.

Post-correction:

Monsters and mythical creatures from around the world are coming to life and joining the parade in a celebration of global folklore and fantasy:\n\nGiant dragon puppet winding through the streets\n- Floating unicorn island with sparkling waterfall mane\n- Phoenix rising from fiery nest on a float\n- Kraken tentacles emerging from a misty sea base\n- Anansi the spider weaving a giant, glowing web\n- Mounted warriors on griffins soaring above the crowd\n- Baku float consuming nightmares with dreamy, swirling colors\n- Qilin float stepping gracefully on clouds of smoke\n- A Jörmungandr float encircling the entire parade route\n- Sphinx float with riddles projected onto its body

## **Appendix J LLM-as-a-Judge Prompts**

We present here the detailed prompts used for relevance and diversity evaluations in Sections 4.3. Both diversity and relevance are assessed using GPT-4o.

**Relevance:** Relevance is assessed using 20 randomly sampled responses per model-prompt-method combination on a three-point scale: irrelevant (0 point), partially relevant (1 point), and relevant (1 point). Both relevant and partially relevant responses are counted as useful outputs when computing the utility rate.

The prompt we used is the following, in which scale’s order is shuffled to avoid position bias. “user prompt” is the prompt used to generate the responses. “response” is the chosen response to be evaluated.

You are an AI assistant tasked with evaluating the relevance of a provided passage to a given user prompt.

Provide your reasoning and classify the passage as “scale[0]”, “scale[1]”, or “scale[2]”.

User prompt: user prompt

Passage to evaluate: response

**Diversity:** Diversity is assessed using 20 randomly sampled response pairs per model-prompt-method combination on a three-point scale: almost identical (0 point), partially similar (1 point), and mostly different (2 points). The normalized mean score represents the method’s diversity level.

The prompt we used is the following, in which scale’s order is shuffled to avoid position bias. “user prompt” is the prompt used to generate the responses. “response0” and “response1” are the chosen pairs to be compared.

You are an AI assistant. Your task is to evaluate the similarity between two passages based on the user prompt provided. Carefully consider and compare the following aspects: 1) Concepts presented, 2) Writing style, 3) Tone of voice, 4) Perspectives, and 5) Opinions.

- For creative writing, pay close attention to the story line. If they are different, then classify as “Mostly Different”.

- For argumentative essay, pay close attention to the arguments, logic and examples used. If these elements are different, then classify as “Mostly Different”.

- For history and science questions, pay close attention to the concepts, perspectives, opinions, and the tone used.

If these elements are different, then classify as “Mostly Different”. After analyzing, provide a brief explanation of your reasoning. Then, classify the passages into one of these categories: “{scale[0]}”, “{scale[1]}”, “{scale[2]}”,

User prompt: {user prompt}

Passage 1: {response0}

Passage 2: {response1}

## Appendix K Human Evaluation of Relevance Judge

To validate the alignment between our LLM-based relevance judge and human judgments, we sampled 500 responses (stratified across methods and models) and collected human annotations via Amazon Mechanical Turk. Each response was independently evaluated by three raters on a three-level scale (Relevant, Partially Relevant, Irrelevant). To ensure

annotation quality, we randomly injected attention-test questions comprising 25% of each survey. Raters who failed to achieve 100% on these attention tests were discarded.

Among the 108 responses the LLM judge rated as Partially Relevant, pairwise inter-rater agreement was only 73%, compared to 90% for the 381 responses rated as Relevant. When collapsing these two categories into a single Relevant class, the binary distinction increases the majority-vote agreement with the LLM judge increased to 97% on average, with inter-annotator agreement Fleiss'  $\kappa = 0.50$  (moderate agreement).

For our research question, the meaningful distinction is whether outputs are useful for search quests (relevant or partially relevant) versus unusable (irrelevant). On this binary classification, 97% agreement exceeds agreement rates reported in recent LLM evaluation studies: Zheng et al. [21] report that GPT-4 achieves 85% agreement with human experts on MT-Bench (exceeding the 81% human-human agreement rate) and 83–87% agreement on Chatbot Arena. In domain-specific search query parsing tasks, LLM judges have demonstrated approximately 90% agreement with human judgments [1].

## References

- [1] Mehmet Selman Baysan, Serkan Uysal, Irem Islek, Cagla Cig Karaman, and Tunga Gungor. Llm-as-a-judge: automated evaluation of search query parsing using large language models. *Frontiers in Big Data*, 8, 2025.
- [2] William G Cochran. *Sampling Techniques*. John Wiley & Sons, 3rd edition, 1977.
- [3] Educational Testing Service. Pool of analytical writing topics. <https://www.ets.org/pdfs/gre/analytical-writing-pool.pdf>, 2024. GRE General Test Resource.
- [4] Gryphe. Chatgpt-4o writing prompts dataset. <https://huggingface.co/datasets/Gryphe/ChatGPT-4o-Writing-Prompts>, 2025. Accessed: 2026-01-31.
- [5] k mktr. improved-flux-prompts dataset. <https://huggingface.co/datasets/k-mktr/improved-flux-prompts>, 2024.
- [6] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*, 2024.
- [7] Morgane Laouenan, Palaash Bhargava, Jean-Benoît Eyméoud, Olivier Gergaud, Guillaume Plique, and Etienne Wasmer. A cross-verified database of notable people, 3500bc-2018ad. *Scientific Data*, 9(1):290, 2022.
- [8] Jiwei Li, Will Monroe, and Dan Jurafsky. A simple, fast diverse decoding algorithm for neural generation, 2016. <https://arxiv.org/abs/1611.08562>.
- [9] Chris Yuhao Liu, Liang Zeng, Jiakai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*, 2024.
- [10] Queenie Luo, Michael J. Puett, and Michael D. Smith. A perspectival mirror of the elephant. *Commun. ACM*, 67(8):98–105, August 2024. <https://doi.org/10.1145/3670241>.
- [11] Clara Meister, Tiago Pimentel, Luca Malagutti, Ethan Wilcox, and Ryan Cotterell. On the efficacy of sampling adapters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1437–1455, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [12] Joan Meyers-Levy and Alice M Tybout. Schema congruity as a basis for product evaluation. *Journal of Consumer Research*, 16(1):39–54, 1989.
- [13] Pavithree. Askhistorians dataset. <https://huggingface.co/datasets/Pavithree/askHistorians>, 2024.

- [14] Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. Is temperature the creativity parameter of large language models?, 2024.
- [15] Rolf Reber, Norbert Schwarz, and Piotr Winkielman. Processing fluency and aesthetic pleasure: Is beauty in the perceiver’s processing experience? *Personality and Social Psychology Review*, 8(4):364–382, 2004.
- [16] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Miresghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.
- [17] Amos Tversky and Daniel Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2):207–232, 1973.
- [18] Robert B Zajonc. Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, 9(2p2):1–27, 1968.
- [19] Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R. Tomz, Christopher D. Manning, and Weiyan Shi. Verbalized sampling: How to mitigate mode collapse and unlock llm diversity, 2025. <https://arxiv.org/abs/2510.01171>.
- [20] Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. Noveltybench: Evaluating language models for humanlike diversity, 2025. <https://arxiv.org/abs/2504.05228>.
- [21] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. <https://arxiv.org/abs/2306.05685>.