Continuous Context Sampling Allows Extending Diversity Boundaries of Large Language Models

Mateusz Bystroński[1] **Doheon Han**[2] **Nitesh V. Chawla**[2] **Tomasz Kajdanowicz**[1]

[1]Wrocław University of Science and Technology
[2]University of Notre Dame

Correspondence: mateusz.bystronski@pwr.edu.pl

Abstract

Starting from the observation that conditioning a poetry-writing prompt with a pancake recipe leads an LLM to produce a coherent poem incorporating pancake-related content and, more broadly, that such contexts arrange themselves into a structured semantic vector space, we argue that this renders the space explorable. By sampling it and using the resulting continuous representations to condition an LLM's generation distribution, we can systematically expand the model's reachable semantic range. We introduce a framework that requires no modification of LLM parameters and operationalizes this idea by conditioning LLM's generation via an xRAG-style projector (Cheng et al., 2024). Our experiments demonstrate that this sampling-based conditioning substantially increases generative diversity, with direct benefits for enhancing divergent thinking, a core facet of creativity, in language models.[1]

1 Introduction

Large language models (LLMs) have become the foundation of modern NLP systems, yet their generative behavior exhibits a persistent limitation: despite using stochastic decoding methods such as temperature or nucleus sampling, repeated generations from the same prompt tend to be semantically similar. This lack of variance constrains applications that rely on broad exploration of the semantic space, including synthetic data generation, brainstorming, and other tasks requiring divergent thinking capabilities, like hypothesis generation.

A widely adopted strategy for increasing diversity is to manipulate the *context* presented to the model, for example through paraphrasing, persona shifts, stylistic changes, or multi-agent discussions. Although effective to a degree, these methods operate over a finite (or effectively finite) set of contexts. Because the conditional distribution $p_\theta(y \mid c)$ associated with each context is known to exhibit low variance (Zhang et al., 2025), the diversity obtainable by marginalizing over such a finite set is inherently limited. Empirically, this manifests as rapid saturation: after only a handful of samples, prompt- and agent-based methods cease to discover new semantic variants.

In this work, we propose a different perspective. Instead of relying solely on symbolic prompt manipulations, we introduce a *continuous* conditioning variable in the model's semantic space. For a given input, we derive a latent representation and modulate the generation context via a multimodal projector, following the xRAG mechanism (Cheng et al., 2024). Crucially, this conditioning operates directly in the token-embedding space of the LLM and therefore requires *no fine-tuning* of the underlying model. By sampling this continuous manifold, we enable the model to access semantic variations that are unattainable through prompt engineering alone.

A central challenge is determining how to sample the latent variable. We show that classical latent models such as VAEs (Kingma and Welling, 2014) are ill-suited for this task due to a topological mismatch between their unimodal latent priors and the clustered, multi-component structure of LLM semantic representations (Cai et al., 2021). A further difficulty is that it is not a priori clear which regions of the latent space carry semantically valuable conditioning signal. We argue, and empirically demonstrate (see Section 5.2, Appendix A) that this choice is largely arbitrary: any context can carry useful conditioning signal, and moreover the xRAG projector appears robust to out-of-distribution latent inputs, producing coherent generations across a broad range of sampled points.

Our experimental results demonstrate that continuous semantic sampling substantially increases the variance of generated outputs without compromising quality. On the NOVELTYBENCH benchmark,

[1]Source code is available at github.com

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 1436–1450

our method uncovers new semantic classes even at large sampling budgets while maintaining high utility. On the Alternative Uses Test (AUT), a classical measure of divergent thinking, latent-space exploration yields the highest originality scores across all settings, approaching the practical upper bound of the scoring scale.

Contributions. This work makes the following contributions:

- We introduce a plug-in latent-conditioning framework that modulates a LLM distribution through sampling in semantic space, requiring no modification of model parameters.
- We provide a topological analysis explaining why VAE-style latent methods cannot align with the clustered geometry of LLM semantic activations.
- We empirically demonstrate substantial gains in semantic diversity and divergent thinking performance on NOVELTYBENCH and AUT task.

2 Related Work

A growing line of work has argued that contemporary LLMs suffer from mode collapse and limited semantic variability despite stochastic decoding. NoveltyBench (Zhang et al., 2025) introduces a benchmark and metric suite specifically designed to assess the ability of models to produce multiple distinct and high-quality responses to a single prompt. Instead of relying on surface-level overlap, it clusters outputs into abstract equivalence classes and reports diversity in terms of the number of occupied classes and their utility. Our work adopts this abstraction-based view of diversity and builds on NoveltyBench as a primary evaluation environment.

Classical approaches increase variability by modifying the decoding procedure, e.g., through temperature scaling or nucleus sampling. These methods flatten the output distribution but do not exploit structure across multiple generations and often exhibit a sharp diversity–quality trade-off. Inference-time methods based on diverse beam search (Cho, 2016; Li and Jurafsky, 2016; Vijayakumar et al., 2017; Kulikov et al., 2019) and related decoding heuristics similarly operate on the token distribution of a *fixed* conditional $p_\theta(y \mid c)$: they ensure that hypotheses in a beam differ lexically, but they do not explicitly reason about semantic redundancy between complete responses. Empirical comparisons with simple temperature tuning suggest that these decoding tweaks only partially alleviate diversity collapse and can harm quality when pushed too far (Ippolito et al., 2019; Zhang et al., 2021; Peeperkorn et al., 2024; Shur-Ofry et al., 2024).

Recent work proposes more principled training-time mechanisms. Early approaches encourage diversity by modifying the maximum-likelihood objective itself: mutual-information objectives discourage generic replies (Li and Jurafsky, 2016; Li et al., 2016), unlikelihood losses penalize degenerate loops and repetitions (Welleck et al., 2020), and smoothing or reshaping the target distribution (e.g., data-dependent Gaussian priors or explicitly diffuse targets) biases models toward broader output distributions (Li et al., 2020b; Zhang et al., 2024). More recent preference-based methods encode diversity directly into the reward or preference model: Diverse Preference Optimization and related objectives (Lanchantin et al., 2025; Slocum et al., 2025) and Creative Preference Optimization (Ismayilzada et al., 2025) jointly optimize for quality and variety of generations. In the context of reasoning models, online RL methods further adjust rewards or weighting schemes to encourage exploration of diverse solution trajectories (Cui et al., 2025; Cheng et al., 2025; Liu et al., 2025; Zeng et al., 2025; Kirk et al., 2024). While effective, all of these techniques require supervised fine-tuning or RL-style updates to the base model, which we explicitly avoid: our goal is to increase diversity *without* modifying LLM parameters. Our method is therefore complementary to these approaches.

A complementary direction explores post-hoc guidance during generation. G2 (Ruan et al., 2025) (Guided Generation) uses an auxiliary classifier to steer the model towards more diffuse response distributions while maintaining task usefulness, and serves as a strong decoding-based baseline in our experiments. Our approach is orthogonal: rather than shaping token probabilities via an external guidance signal, we modify the semantic *conditioning* itself by sampling it from a semantic space.

Beyond generic methods, some work targets diversity in application-specific formats. For instance, Holysz et. al. have explored JSON-based prompting schemes to induce structurally diverse outputs in medical scenario (Holysz et al., 2025); however, such JSON schemas are highly task-specific and do not naturally generalize to open-

ended semantic variation. In contrast, our method operates at the level of continuous text embeddings and applies uniformly across tasks.

A long-standing line of research uses variational autoencoders to endow language models with a continuous latent code controlling generation. OPTIMUS (Li et al., 2020a) and follow-up work (Zhang et al., 2023) train VAEs on top of large pretrained models to organize sentences in a latent space that supports interpolation, traversal, and conditional control. These methods, however, require optimizing the decoder. This introduces the usual costs and risks of model fine-tuning, including potential catastrophic forgetting of pretrained semantics. In Appendix B we deep dive into this phenomenon.

xRAG (Cheng et al., 2024) demonstrates that LLMs can be conditioned directly through dense semantic vectors injected via a multimodal projection layer. Their goal, however, is orthogonal to ours: xRAG uses embedding-based conditioning to *compress external documents* for efficient RAG, keeping the model aligned to retrieved evidence. We instead generalize this mechanism to modulate the model's *internal* semantic state, treating the conditioning vector not as compressed context but as a latent variable for controlled exploration. Thus, while xRAG establishes that continuous conditioning is feasible without fine-tuning, our work leverages this capability to expand semantic variance in generation.

Our second line of evaluation concerns divergent thinking and creativity. Recent work has begun to systematically assess language creativity of LLMs and humans using batteries of psychological tests. Dinu et al. propose an integrated creativity suite (Dinu and Florescu, 2025), including the Alternative Uses Test (AUT), and report that strong LLMs can approach or slightly surpass human performance under certain conditions. These works focus primarily on measuring creativity, not on algorithmic mechanisms for increasing it. We adopt AUT as an evaluation task and interface with an existing automatic originality scoring framework (Organisciak et al., 2023), which introduced a method for automated scoring that demonstrates high alignment with human annotators.

Multi-agent schemes have been explored as a way to enhance creativity and diversity by simulating human-like group discussions. Lu et al. propose *LLM Discussion* (Lu et al., 2024), a three-phase multi-agent, role-play framework which significantly improves performance on AUT and other creativity tests compared to single-agent baselines and simpler multi-agent setups. Subsequent surveys further document the promise of LLM-based multi-agent systems for creativity. Multi-agent schemes have been explored as a way to enhance creativity and diversity by simulating human-like group discussions. Lu et al. propose *LLM Discussion* (Lu et al., 2024), a three-phase multi-agent, role-play framework which significantly improves performance on AUT and other creativity tests compared to single-agent baselines and simpler multi-agent setups. Subsequent surveys further document the promise of LLM-based multi-agent systems for creativity. Conceptually, such methods still operate within the symbolic prompt-based paradigm: since agent context has a linear order, it translates into in-context regeneration (Zhang et al., 2025) which is shown to exhibit low variance. Our experiments corroborate this: increasing the depth of LLM Discussion yields only marginal gains before diversity saturates. Our experiments corroborate this: increasing the depth of LLM Discussion yields only marginal gains before diversity saturates.

A related line of research studies interpolation-based mechanisms for extending datasets and improving coverage of underrepresented regions in feature space. Classical oversampling method SMOTE (Chawla et al., 2002) generate synthetic examples by linear interpolation between nearest neighbors and have been widely and successfully applied to *tabular data*. Deep variants such as DeepSMOTE (Dablain et al., 2023) extend this idea to learned representation spaces, enabling interpolation in latent embeddings learned by neural encoders in vision domain. Our approach is inspired by these techniques, but extends the paradigm to natural language domain.

3 Method

We begin with the observation (see Appendix A) that conditioning an LLM on an additional context, even one that is seemingly unrelated to the original prompt, can lead to more original outputs. However, such contexts are difficult to sample directly in text space: textual prompts are discrete, highly structured, and hard to perturb in a controlled way. We therefore move to the embedding space.

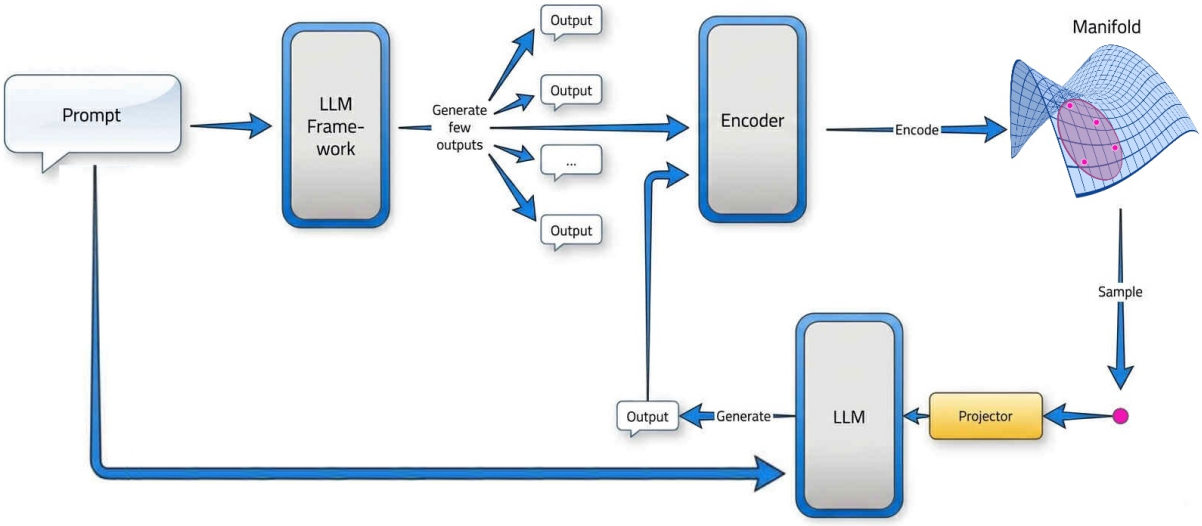

Figure 1: **Continuous context sampling.** Given an input prompt, a stochastic latent vector z is drawn from a sampling distribution $q_\phi(z \mid e)$ and mapped through an xRAG-style projector (Cheng et al., 2024) into the LLM's input-embedding space. The LLM generates a candidate output conditioned on this embedding; a constraint-preserving realignment step then rewrites the candidate so that it satisfies the original task specification. We propose two distributions for z: random isotropic sampling at the projector's natural input scale, and anchor-based interpolation over a small seed set A_x.

3.1 Continuous Context Sampling

We add a continuous stochastic conditioning channel to a frozen LLM. For an input x, an encoder

$$e = E(x) \in \mathbb{R}^d \tag{1}$$

produces a dense semantic representation. A latent variable

$$z \sim q_\phi(z \mid e), \qquad z \in \mathbb{R}^d \tag{2}$$

is drawn from a sampling distribution q_ϕ, whose support is a continuous subset of $\mathbb{R}^d$. The latent modulates context construction:

$$c = g(x, z), \tag{3}$$

where g injects z into the LLM's input-token embedding space via the multimodal projector of (Cheng et al., 2024). Because g acts in embedding space and the LLM is frozen, the procedure is fully plug-in. The LLM then generates

$$Y \sim p_\theta(\cdot \mid c) = p_\theta(\cdot \mid g(x, z)), \tag{4}$$

and the marginal over outputs becomes

$$p(y \mid x) = \int p_\theta\big(y \mid g(x, z)\big)\, q_\phi\big(z \mid E(x)\big)\, dz. \tag{5}$$

For any feature functional $f(Y)$, the law of total variance gives

$$\begin{aligned}\mathrm{Var}[f(Y) \mid x] = \mathbb{E}_z\big[\,\mathrm{Var}[f(Y) \mid x, z]\big] + \\ \mathrm{Var}_z\big(\mathbb{E}[f(Y) \mid x, z]\big).\end{aligned} \tag{6}$$

Continuous latent conditioning allows z to vary smoothly across a high-dimensional region, so that $\mathrm{Var}_z\big(\mathbb{E}[f(Y) \mid x, z]\big)$ can be substantially larger.

3.2 Sampling Distribution

We explore two complementary instantiations of $q_\phi(z \mid e)$. Both operate in the encoder's output space (dimension d) and are passed through the same projector. Both are evaluated in Section 4.

(i) Random isotropic sampling. The simplest choice. We draw

$$z = e + \sigma\, \eta, \qquad \eta \sim \mathcal{N}(0, I), \tag{7}$$

where σ is calibrated to the natural per-coordinate standard deviation of the encoder's output distribution. We measured this quantity for the encoder used in our experiments (Section 4) and obtained $\sigma \approx 4.77$ as a reference scale. Importantly, downstream sensitivity to σ is mild (see ablation in Section 5.2). Operationally, the random sampler does not require careful tuning per task or domain.

(ii) Anchor-based interpolation. A more interpretable sampler. We collect a small set of seed generations $y_1, \ldots, y_m$ for prompt x (using any base method — for instance, decoding from the LLM with varied seeds, or seeds from an auxiliary controller such as G2 (Ruan et al., 2025)), embed them via $E(\cdot)$ to obtain anchors

$$A_x = \{e_1, \ldots, e_m\}, \tag{8}$$

and sample two anchors and an interpolation coefficient

$$(i, j) \sim \pi, \qquad \lambda \sim \rho. \tag{9}$$

The latent is constructed as

$$z = (1 - \lambda)\, e_i + \lambda\, e_j. \tag{10}$$

We restrict ourselves to interpolation-based families because the encoder used in our experiments is contrastively-trained, which tends to organize semantic classes into approximately convex clusters (Cai et al., 2021): linear combinations of in-cluster embeddings yield stable variations, and large λ produces extrapolative samples that leave the seed cluster while remaining in the encoder's natural output range. Concretely, in our main experiments we use $\lambda \sim \mathcal{U}([6, 10] \cup [-6, -10])$, chosen based on ablation as best performing λ values (see Section 5.2).

3.3 Constraint-Preserving Realignment

Conditioning on z widens the effective context $c = g(x, z)$ but, in constraint-sensitive tasks, can introduce deviations from the original specification (e.g. format, item count, role). We therefore define a realignment operator

$$\tilde{Y} = r(Y, x), \qquad Y \sim p_\theta(\cdot \mid g(x, z)), \tag{11}$$

implemented as a brief follow-up call to the same LLM with a terse rewrite prompt that asks for an output satisfying the original specification while preserving the candidate's content. Because r is applied *after* latent conditioning, the variance decomposition becomes

$$\begin{aligned}\mathrm{Var}[f(\tilde{Y}) \mid x] = \mathbb{E}_z\Big[\,\mathrm{Var}_Y\big[f(r(Y, x))\big]\Big] \\ + \mathrm{Var}_z\Big(\mathbb{E}_Y\big[f(r(Y, x))\big]\Big),\end{aligned} \tag{12}$$

and the second term remains nonzero whenever the distribution of realigned outputs $r(Y, x)$ varies with z.. In other words, realignment can recover task adherence without erasing latent-induced semantic variation. We verify both sides of this claim empirically in Section 5.1 (the realignment step contributes a consistent utility lift; the diversity ranking is preserved when it is disabled).

3.4 Why not a VAE?

A natural alternative is to obtain $q_\phi(z \mid e)$ from a variational autoencoder, which provides a smooth latent manifold with a known prior. Prior work indicates that VAEs in this setting require an *unfrozen* decoder to align the latent space with the LLM's semantic geometry (Li et al., 2020a; Zhang et al., 2023), which contradicts our plug-in goal. Appendix B provides a fuller discussion. The two samplers we propose avoid this issue entirely by sampling in the encoder's existing output space and relying on the pretrained projector for alignment.

4 Experiments

In all experiments we use **Mistral-7B-Instruct** (Jiang et al., 2023) as the base language model. Semantic representations are obtained from the **SFR-Embedding-Mistral** encoder (Meng et al., 2024); the projector g is the publicly released xRAG projector (Cheng et al., 2024) trained on this encoder–LM pair.

4.1 Generation Diversity (NoveltyBench)

We evaluate on the *curated* subset of NOVELTY-BENCH (Zhang et al., 2025), which measures generation diversity using the *Distinct* metric, i.e. the number of abstract equivalence classes covered by a set of generations:

$$\mathrm{distinct}_k := \big|\{c_i : i = 1, \ldots, k\}\big|,$$

where c_i denotes the equivalence class assigned to the i-th generation by the benchmark classifier. Since diversity can be increased by producing uninformative or misaligned outputs, we also report the benchmark's *Utility* metric:

$$\mathrm{utility}_k := \frac{1 - p}{1 - p^k} \sum_{i=1}^{k} p^{i-1} \cdot \mathbf{1}\Big[c_i \neq c_j \; \forall j < i\Big] \cdot u_i,$$

with patience $p = 0.8$, following (Zhang et al., 2025).

We evaluate both metrics across multiple generation budgets $k \in \{10, 15, 20, 25, 30\}$. It is particularly important for diversity to increase with larger sampling budgets, especially in applications such

as synthetic data generation, where the marginal gains from additional samples directly translate into broader and more representative coverage of the underlying semantic space.

We compare three groups of methods. First, we report prompt- and decoding-time baselines: Standard (Zhang et al., 2025), In-context (Zhang et al., 2025), and G2 (Ruan et al., 2025). Second, we evaluate our method in a standalone configuration using random latent sampling without any seed generations (OURS, RANDOM, NO SEEDS). Third, we evaluate hybrid variants in which G2 produces the initial seed generations, after which our latent sampler generates the remaining outputs. In the hybrid setting, G2 provides the first 30% of generations for $k \in \{10, 15, 20\}$ and the first 20% for $k \in \{25, 30\}$. We test both random latent sampling and anchor-based interpolation. Sampling parameters are selected according to the ablations in Section 5.2: $\sigma = 4.77$ for random sampling and $\lambda \sim \mathcal{U}([6, 10] \cup [-6, -10])$ for interpolation. All configurations use the same encoder, projector, base model, decoding settings, seed, and realignment prompt.

Table 1 shows that the standalone version of our method consistently increases diversity relative to the baselines. At $k = 30$, it achieves $D = 16.50$, compared to 13.60 for G2 and 13.31 for the in-context baseline. This comes at a cost in Utility relative to G2, however, its Utility remains above the in-context baseline across all budgets.

The hybrid variants provide the strongest overall results. They improve over G2 on both Distinct and Utility at every evaluated budget, indicating that latent-space exploration is complementary to G2. In particular, the hybrid random sampler reaches the highest diversity at all budgets, with $D = 17.77$ at $k = 30$, while maintaining the highest utility. This indicates that our method is especially effective in preserving long-term utility as the generation budget increases.

4.2 Impact of Diverse Generation on Divergent Thinking Capabilities

We examine how increased generation diversity translates into improved divergent thinking capabilities. We laveraged psychological Alternative Uses Test (AUT), a classical psychological assessment of divergent thinking and creative potential (Lu et al., 2024; Silvia et al., 2008; Dinu and Florescu, 2025). Participants in an AUT task are asked to propose unusual and non-obvious uses for everyday objects. Responses are traditionally evaluated for originality, flexibility, and fluency. We focus on core metric, originality (Lu et al., 2024).

We used the originality scoring framework from (Organisciak et al., 2023), which provides automated originality ratings (from 1 to 5) aligned with human-labeled AUT datasets, using model **ocsai-4o**, which is said by authors to be good for English Alternate Uses scoring. Following the creativity literature (Silvia et al., 2008; Dinu and Florescu, 2025), which recommends focusing on only a few ideas, we report:

- **Top–1 originality**: the most original idea,
- **Top–2 originality**: the mean of the two most original ideas,
- **Top–3 originality**: the mean of the three most original ideas.

We laveraged dataset from (Lu et al., 2024) For each AUT prompt, we used the output of the multi-turn **LLM discussion** as the anchor set. Previous work reports that this discussion-based method does not scale (Lu et al., 2024); we confirmed this in our own setup by running discussions of varying lengths. The best-performing discussion depth was selected, and its output served as both our baseline and our anchor points.

We evaluated G2 using the first generations from the LLM discussion as contextual seeds. Our proposed method was likewise evaluated using the LLM discussion outputs as anchors. For both methods, we ran 500 generations per approach. We did not apply any alignment step during the AUT experiment. Since the task focuses purely on the originality of semantic content rather than consistency of style or structure, additional stylization mechanisms were unnecessary and were therefore omitted.

The results presented in Table 2 highlight three key observations. First, increasing the depth of multi-agent LLM discussion yields only marginal gains: Top-1 plateaus after a few rounds, confirming that this method does not scale. Second, expanding the diversity of generations has a direct and measurable impact on creativity. As illustrated at Figure 2, Top–1, Top–2, and Top–3 scores steadily improve as more latent samples are drawn, indicating that broader exploration of the semantic space translates into consistently more original ideas. Finally, our latent-space method achieves

Method	Metric	NoveltyBench (k generations)				
		$k=10$	$k=15$	$k=20$	$k=25$	$k=30$
Standard	D	4.37	4.64	5.43	6.20	6.79
	U	3.62	0.84	0.82	0.78	0.79
In-context	D	7.13	9.35	11.70	12.68	13.31
	U	2.97	2.81	2.71	2.76	2.71
G2 (Ruan et al., 2025)	D	6.21	8.29	10.27	12.04	13.60
	U	4.52	4.48	4.31	4.33	4.33
Ours, random, no seeds	D	7.09	9.68	11.62	13.71	16.50
	U	3.77	3.55	3.45	3.48	3.73
Hybrid: G2 + interp	D	7.44	10.08	12.26	14.26	16.18
	U	4.95	**4.77**	4.69	**4.69**	**4.68**
Hybrid: G2 + random	D	**7.68**	**10.67**	**12.95**	**15.30**	**17.77**
	U	**4.96**	**4.77**	**4.70**	4.65	4.66

Table 1: NoveltyBench, curated subset. Distinct (D) and Utility (U) across generation budgets. The standalone latent sampler achieves the highest diversity among non-hybrid methods, while the hybrid variants obtain the strongest overall results by combining G2's high-utility seed generations with additional latent-space exploration.

Method	**Top–1**	**Top–2**	**Top–3**
LLM Discussion (1 round)	4.17	4.06	3.95
LLM Discussion (3 rounds)	4.57	4.52	4.49
LLM Discussion (5 rounds)	4.58	4.55	4.50
LLM Discussion (7 rounds)	4.58	4.56	4.53
G2	4.93	4.92	4.90
Ours	**4.99**	**4.98**	**4.95**

Table 2: Comparison of AUT originality scores across discussion-based baselines, G2, and our latent-space exploration method. Scores are reported as Top–1, Top–2, and Top–3 originality.

the strongest originality across all evaluation settings, reaching a Top–1 score of 4.99. This value is extremely close to the practical upper bound of the AUT scale - 5, effectively demonstrating that our method pushes the model's creative capacity to the limits.

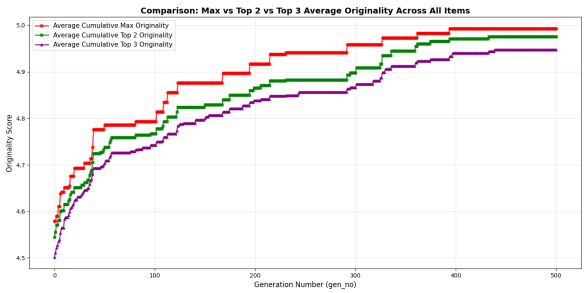

Figure 2: Cumulative originality curves for our latent-space exploration method. As more latent samples are drawn, the Top–1, Top–2, and Top–3 originality scores steadily increase.

5 Ablation Studies

We use ablations to isolate the main components of the method: stochastic latent sampling, the projector, and the realignment step. Unless otherwise stated, all experiments use the curated NoveltyBench subset with $k = 15$, the same encoder, projector, base model, decoding settings, and seed. Most ablations are run in the hybrid regime, where G2 produces the first 30% of generations and the remaining outputs are generated by our latent sampler. This setting provides a common anchor pool for all anchor-based variants.

5.1 Sampling Strategy and Realignment

We first vary only the way the conditioning vector z is selected, keeping the projector, anchor set, and decoder fixed. We compare deterministic anchor representatives — SINGLE, MEAN, and MEDOID — with stochastic variants: a tiny Gaussian perturbation around an anchor (GAUSS), anchor-based interpolation (INTERP), and random isotropic sampling (RANDOM). Each variant is evaluated with and without realignment.

Table 3 shows that deterministic conditioning collapses to roughly four distinct classes, regardless of whether the selected point is a single anchor, the centroid, or the medoid. Diversity increases only when the conditioning vector is sampled stochastically. A very small Gaussian perturbation already improves coverage, but the strongest results are obtained by high-spread sampling: INTERP reaches $D = 10.08$, while RANDOM reaches $D = 10.67$.

The realignment step is not responsible for the

	Realign. ON		*Realign. OFF*	
Sampling strategy	D (#)	U	D (#)	U
SINGLE	4.20	3.86	3.96	3.74
MEAN	4.19	3.82	4.06	3.79
MEDOID	4.18	3.84	3.99	3.73
GAUSS (σ=0.05)	6.74	4.38	6.11	4.18
INTERP	10.08	**4.77**	9.83	4.15
RANDOM (σ=4.77)	**10.67**	**4.77**	**10.13**	4.00

Table 3: Sampling-strategy ablation. All rows share the same projector and anchor set; only the choice of $q_\phi(z \mid e)$ changes. Distinct (D) counts equivalence classes out of 15 generations, while Utility (U) measures discounted task utility. Realignment is toggled in the right block.

diversity gain. Without realignment, both INTERP and RANDOM still substantially outperform all deterministic baselines with realignment. Its main effect is on Utility, while stochastic latent conditioning expands semantic coverage.

5.2 Latent Spread is the Main Lever

We next compare the two high-spread samplers from Section 3.2: anchor-based interpolation and random isotropic sampling. Although they differ in parametric form, both are controlled by a spread parameter: λ for interpolation and σ for random sampling. We sweep these parameters to test whether performance depends on the sampler geometry or on moving far enough from the seed cluster.

Table 4 shows a threshold effect. When the spread is too small, diversity drops, once the spread is large enough, both samplers enter a stable high-diversity regime. For INTERP, this occurs around $\lambda \geq 6$; for RANDOM, it already occurs by $\sigma \geq 0.5$.

Above this threshold, the exact sampler form is secondary. Random isotropic sampling is consistently about one distinct class higher than interpolation, but both methods preserve essentially the same Utility. The random sampler is also remarkably stable: Distinct remains within a narrow band across more than four orders of magnitude in σ. We do not characterize the precise mechanism behind this insensitivity here; two natural hypotheses are that the projection mechanism is robust to input norm (e.g., via downstream normalization in the decoder), or that the trained projector pulls out-of-distribution latents back onto a region of its output manifold to which the LLM is already conditioned. Understanding this mechanism is left to future work. We use $\sigma = 4.77$ and $\lambda \sim \mathcal{U}([6, 10])$ in the remaining experiments.

Configuration	D (#)	U
*Anchor-based (*INTERP*), λ sweep*		
$\lambda = 0.5$	6.30	4.27
$\lambda = 2$	8.34	4.63
$\lambda = 4$	9.05	4.69
$\lambda = 6$	9.55	**4.78**
$\lambda = 8$	9.66	4.73
$\lambda = 10$	9.62	4.77
Random isotropic, σ sweep (σ_{nat}=4.77)		
$\sigma = 0.05$	7.75	4.59
$\sigma = 0.5$	10.57	4.76
$\sigma = 1$	**10.82**	4.75
$\sigma = 2$	10.65	4.76
$\sigma = 4.77$	10.67	4.77
$\sigma = 20$	10.71	4.76
$\sigma = 50$	10.71	4.75
$\sigma = 500$	10.76	4.75
$\sigma = 5{,}000$	10.63	4.72
$\sigma = 10{,}000$	10.75	4.71

Table 4: Spread sweep, $k = 15$, NoveltyBench curated subset. INTERP reaches a plateau around $D \approx 9.6$ and $U \approx 4.77$ once $\lambda \geq 6$. RANDOM reaches a plateau around $D \approx 10.7$ and $U \approx 4.75$ once $\sigma \geq 0.5$, and remains stable across more than four orders of magnitude in σ.

5.3 Anchor Quality and Number of Seeds

We then test whether the anchor-based sampler depends strongly on the quality or number of seed generations. We perturb anchors with Gaussian noise and separately cap the number of available seeds.

Table 5 shows that anchor identity is not a fragile component of the method. Even when anchors are perturbed with $\sigma = 5$, reducing their cosine similarity to the original anchors to approximately 0.69, performance remains close to the unperturbed setting ($D = 10.16$, $U = 4.74$). Reducing the number of seeds has a larger effect on Utility, and modest on Diversity.

5.4 Sampling Must Occur in the Projector Input Space

The proposed samplers draw z in the encoder output space and then map it through the learned projector into the LLM input-embedding space. We test whether this projector is necessary by bypassing it: instead of sampling z and using $g(z)$, we sample a vector z' directly in the LLM input-embedding space and insert it at the <xRAG> position.

We sweep σ_{LM} across three scales: the natural per-coordinate standard deviation of Mistral token embeddings ($\sigma_{\text{LM}} = 0.0027$), an intermediate scale ($\sigma_{\text{LM}} = 1.0$), and a scale matched to the

Configuration	D (#)	U
INTERP (full anchors, no noise)	10.08	**4.77**
Anchor noise σ (cos to original shown for reference)		
$\sigma = 0.10$ (cos ≈ 1.00)	10.00	4.76
$\sigma = 0.25$ (cos ≈ 0.99)	9.80	4.69
$\sigma = 1$ (cos ≈ 0.98)	**10.28**	4.75
$\sigma = 2$ (cos ≈ 0.92)	9.98	4.74
$\sigma = 5$ (cos ≈ 0.69)	10.16	4.74
Seed count k		
$k = 1$	8.39	4.11
$k = 2$	8.99	4.46
$k = 2,\ \sigma = 0.10$	9.19	4.43

Table 5: Weak-seed ablation. The anchor sampler is robust to substantial anchor perturbations. Seed scarcity is more disruptive, but even a single seed remains well above deterministic conditioning baselines.

Configuration	D (#)	U
INTERP (full pipeline)	10.08	**4.77**
random (full pipeline, $\sigma = 4.77$)	**10.67**	**4.77**
Bypass projector: sample directly in LLM space		
$\sigma_{\text{LM}} = 0.0027$ (natural)	5.91	4.23
$\sigma_{\text{LM}} = 1.0$ (intermediate)	6.58	4.32
$\sigma_{\text{LM}} = 200$ (projector image)	6.70	4.38

Table 6: Bypass-projector ablation. Sampling directly in the LLM input-embedding space loses roughly four distinct classes relative to the full pipeline, even when the sampling scale is matched to the projector output distribution.

projector output distribution ($\sigma_{\text{LM}} = 200$).

Table 6 shows that bypassing the projector substantially reduces diversity. All direct-LLM-space variants remain near $D \approx 6$–7, whereas the full pipeline reaches $D \approx 10$. The failure of the $\sigma_{\text{LM}} = 200$ condition is particularly informative: matching the marginal scale of projector outputs is not enough. The diversity gain depends on sampling in the projector input space and passing the result through the learned mapping.

5.5 Realignment Prompt Variants

Finally, we ablate the realignment prompt. The tested variants range from verbose rewrite instructions to terse format-preserving prompts. We also vary whether the prompt encourages explanatory rewriting, preserves the candidate output conservatively, or explicitly suppresses meta-language.

Table 7 shows that the realignment prompt primarily affects Utility rather than diversity. Overall, diversity is robust to prompt phrasing. This supports the conclusion that the load-bearing component for diversity is stochastic latent conditioning

Realignment prompt variant	D (#)	U
Prompt 1 (strict concise editor)	10.08	**4.77**
Prompt 2 (explanatory editor)	10.44	3.88
Prompt 3 (polished answer rewrite)	9.97	4.39
Prompt 4 (conservative copy editor)	**10.58**	4.41
Prompt 5 (strict anti-meta rewrite)	10.34	4.58

Table 7: Realignment prompt ablation, $k = 15$, hybrid configuration with the INTERP sampler held fixed. Diversity is relatively stable across rewrite prompts, with all variants falling in the narrow range $D \in [9.97, 10.58]$. Utility is substantially more sensitive to prompt choice, varying by up to 0.89. Prompt 1 achieves the highest Utility while remaining close to the highest-Distinct variant.

rather than the rewrite prompt itself. The realignment step should therefore be understood mainly as a utility-preservation mechanism. The verbatim prompt variants are listed in Appendix C.

6 Discussion

Our experimental results demonstrate that the proposed method introduces substantially greater variation in generated outputs, and that this variance translates into responses that are both more diverse and fully comparable in quality to those produced by baseline approaches.

A key insight is that the latent variable we introduce is not a randomness, it is a variable that has its own context that directly influences the model's response, analogous to the role of retrieved evidence in RAG systems, and this context can be explored geometrically. This makes it possible to apply a wide range of heuristics and metaheuristics to text representations. For example, evolutionary crossbreeding can be naturally expressed as a linear combination of text embeddings.

The AUT experiment illustrates this particularly clearly: we started from outputs generated by a complex agent-based method, and used these responses as anchors and further optimized them through latent-space exploration. In essence, the way we did it in experiment was as an evolutionary strategy, where whole population breeds together.

Taken together, these findings open a new perspective on NLP: tasks traditionally limited by the symbolic and contextual nature of natural language can now be addressed using classical methods from computer science, enabled by the continuous and geometrically structured semantic space

Limitations

Despite the empirical gains reported in this work, our approach has several important limitations.

Our experiments are restricted to a single LLM–encoder–projector stack, mainly because pretrained xRAG-style projectors are not broadly available. Applying the method to other LLMs therefore requires training compatible projectors. Future work should evaluate whether the observed gains transfer across model families, scales, instruction-tuning recipes, and embedding backbones.

Although the projector is essential, our ablations only show its empirical role: bypassing it and sampling directly in the LLM embedding space strongly reduces diversity. We do not yet explain the mechanism that makes the learned mapping robust to high-spread or out-of-distribution latent inputs. Future work should analyze whether this behavior comes from implicit normalization, manifold projection, or alignment with semantically usable regions of the LLM embedding space.

The method increases semantic diversity, but the standalone sampler still shows a diversity–utility trade-off. Future work should develop utility-preserving exploration mechanisms, such as adaptive sampling, learned filtering, reward-guided latent search, or stronger task-aware realignment.

Our sampling strategies are intentionally simple: isotropic Gaussian sampling and linear anchor interpolation. Future work should replace these heuristics with geometry-aware samplers that account for local density, cluster structure, and prompt-specific exploration needs.

Acknowledgments

We are deeply grateful to the reviewers for their careful and constructive feedback, which substantially improved the quality of this work. We also thank Peter Organisciak for his prompt assistance with the OCSAI originality scoring framework.

This work was supported by the AITAX (AI Tax Advisor) project under the grant FENG.02.02-IP.05-0314/23, Action 2.2 FIRST TEAM, European Funds for a Modern Economy Programme 2021–2027 (FENG). Calculations have been carried out in the Wroclaw Centre for Networking and Supercomputing (http://www.wcss.pl) as well as using services of CLARIN-PL.

References

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations*.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. 2025. Reasoning with exploration: An entropy perspective. *Preprint*, arXiv:2506.14758.

Xin Cheng, Xun Wang, Xingxing Zhang, Tao Ge, Si-Qing Chen, Furu Wei, Huishuai Zhang, and Dongyan Zhao. 2024. xrag: Extreme context compression for retrieval-augmented generation with one token. *Preprint*, arXiv:2405.13792.

Kyunghyun Cho. 2016. Noisy parallel approximate decoding for conditional recurrent language model. *Preprint*, arXiv:1605.03835.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, LEI BAI, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. 2025. The entropy mechanism of reinforcement learning for reasoning language models.

Damien Dablain, Bartosz Krawczyk, and Nitesh V. Chawla. 2023. Deepsmote: Fusing deep learning and smote for imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):6390–6404.

Anca Dinu and Andra-Maria Florescu. 2025. Testing language creativity of large language models and humans. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 426–436, Albuquerque, USA. Association for Computational Linguistics.

Mikolaj Holysz, Mateusz Bystronski, Grzegorz Aleksander Piotrowski, Grzegorz Chodak, and Tomasz Kajdanowicz. 2025. A multi-stage llm framework for generating realistic synthetic medical datasets. In *Proceedings of the Americas Conference on Information Systems (AMCIS 2025)*.

Daphne Ippolito, Reno Kriz, João Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762, Florence, Italy. Association for Computational Linguistics.

Mete Ismayilzada, Antonio Laverghetta Jr., Simone A. Luchini, Reet Patel, Antoine Bosselut, Lonneke Van Der Plas, and Roger E. Beaty. 2025. Creative

preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9580–9609, Suzhou, China. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations*.

Ilia Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. 2019. Importance of search and evaluation strategies in neural dialogue modeling. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 76–87, Tokyo, Japan. Association for Computational Linguistics.

Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar Sukhbaatar, and Ilia Kulikov. 2025. Diverse preference optimization. *Preprint*, arXiv:2501.18101.

Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. 2020a. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4678–4699, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li and Dan Jurafsky. 2016. Mutual information and diverse decoding improve neural machine translation. *Preprint*, arXiv:1601.00372.

Zuchao Li, Rui Wang, Kehai Chen, Masso Utiyama, Eiichiro Sumita, Zhuosheng Zhang, and Hai Zhao. 2020b. Data-dependent gaussian prior objective for language generation. In *International Conference on Learning Representations*.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. 2025. ProRL: Prolonged reinforcement learning expands reasoning boundaries in large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung yi Lee, and Shao-Hua Sun. 2024. Llm discussion: Enhancing the creativity of large language models via discussion framework and role-play. *Preprint*, arXiv:2405.06373.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfr-embedding-mistral:enhance text retrieval with transfer learning. Salesforce AI Research Blog.

Peter Organisciak, Selcuk Acar, Denis Dumas, and Kelly Berthiaume. 2023. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, 49:101356.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? In *Proceedings of the 15th International Conference on Computational Creativity, ICCC 2024, Jönköping, Sweden, June 17-21, 2024*, pages 226–235. Association for Computational Creativity (ACC).

Zhiwen Ruan, Yixia Li, Yefeng Liu, Yun Chen, Weihua Luo, Peng Li, Yang Liu, and Guanhua Chen. 2025. G2: Guided generation for enhanced output diversity in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14127–14145, Suzhou, China. Association for Computational Linguistics.

Michal Shur-Ofry, Bar Horowitz-Amsalem, Adir Rahamim, and Yonatan Belinkov. 2024. Growing a tail: Increasing output diversity in large language models.

Paul Silvia, Beate Winterstein, John Willse, Christopher Barona, Joshua Cram, Karl Hess, Jenna Martinez, and Crystal Richard. 2008. Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts*, 2:68–85.

Stewart Slocum, Asher Parker-Sartori, and Dylan Hadfield-Menell. 2025. Diverse preference learning for capabilities and alignment. In *The Thirteenth International Conference on Learning Representations*.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2017. Diverse beam search: Decoding diverse solutions from neural sequence models.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In

International Conference on Learning Representations.

Weihao Zeng, Yuzhen Huang, Lulu Zhao, Yijun Wang, Zifei Shan, and Junxian He. 2025. B-STar: Monitoring and balancing exploration and exploitation in self-taught reasoners. In *The Thirteenth International Conference on Learning Representations*.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2021. Trading off diversity and quality in natural language generation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 25–33, Online. Association for Computational Linguistics.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. 2025. Noveltybench: Evaluating language models for humanlike diversity. *Preprint*, arXiv:2504.05228.

Yiming Zhang, Avi Schwarzschild, Nicholas Carlini, J Zico Kolter, and Daphne Ippolito. 2024. Forcing diffuse distributions out of language models. In *First Conference on Language Modeling*.

Yingji Zhang, Danilo S. Carvalho, Ian Pratt-Hartmann, and André Freitas. 2023. Llamavae: Guiding large language model generation via continuous latent sentence spaces. *Preprint*, arXiv:2312.13208.

A Qualitative case study: arbitrary-text context injection

Setup. As a simple qualitative sanity check, we test whether ordinary textual context, injected directly into the prompt, can shift the semantic content of an LLM's output while preserving the original task constraints. We use vanilla Mistral-7B-Instruct and prepend one of three background contexts to each NoveltyBench prompt: no additional context (vanilla), a short pancake recipe (∼120 words), or a short engineering paragraph about cast-iron pipe burst calculations (∼80 words, containing the terms *cast-iron*, *trench*, *hoop stress*, and *granular bedding*). Each context is introduced as "*Background context (use only if relevant)*".

Prompt: "Write a short love poem with 4 lines."

[vanilla] sample 1/3

In your eyes, a universe unfolds,
A constellation of love, my heart's hold.
Your touch, a symphony, melodies untold,
Together, we dance in an eternal twirl.

[vanilla] sample 2/3

In your eyes, a universe unfolds,
A symphony of love, my heart's bold hold.
Your touch, a melody, soft and untold,
Together, our love, a timeless fold.

[vanilla] sample 3/3

In your eyes, my heart finds a home,
A place where love's sweet melody roams,
Your touch ignites a gentle, radiant flame,
Together, our souls in perfect harmony claim.

[+pancake recipe] sample 1/3

In the warmth of your embrace, I find my home,
Your love, like a classic pancake, softly grown.
Together we savor each moment, sweet and whole,
Maple syrup of passion, our hearts entwined, we've grown.

[+pancake recipe] sample 2/3

In the heart of your kitchen, love blooms,
Where the scent of pancakes and maple beams.
Together we stir, our bond renewed,
A classic dance, in every sweet brood.

[+pancake recipe] sample 3/3

In the heart of your kitchen, a dance begins,
Pancakes and love, in the morning's gentle gleam,
Your touch gentle, as batter meets the pan,
A taste of home, our hearts forever entwined.

[+pipe-engineering paragraph] sample 1/3

Amidst the earth's embrace, our hearts entwine,
A duet of souls in cast-iron bond,
Through layers deep, our love shall shine,
Resilient as iron, soft as wind.

[+pipe-engineering paragraph] sample 2/3

> Amidst the earth's embrace, in trench so deep,
> Two souls entwined, in cast-iron keep.
> Hooped in love, beneath the granular sweep,
> In whispered secrets, our hearts to keep.

[+pipe-engineering paragraph] sample 3/3

> Amidst the earth's embrace, so deep and strong,
> A duet of iron and love, in trench we belong.
> Hooped in promise, walled in trust,
> Bound by the pressures, in our bond, we thrust.

This case study shows that even arbitrary additional context can move the model toward a different semantic region without necessarily breaking instruction adherence. The injected text changes what the model writes about, while the original prompt continues to control the high-level task structure.

B VAE Projector Leads to OOD Risk

Let $Z \subset \mathbb{R}^d$ denote the support of the VAE latent distribution used for sampling, and let $H \subset \mathbb{R}^{d_h}$ denote the support of the LLM's semantic decoder–input activations (e.g., first-layer token embeddings). We model *semantic clusters* in the LLM as the path–connected components of high-density regions in H.

Let p_Z be the latent density (typically Gaussian), and let p_H denote the empirical density of decoder activations. For a threshold $\epsilon > 0$, define the super-level sets

$$Z_\epsilon = \{\, z \in Z : p_Z(z) \geq \epsilon \,\},$$

$$H_\epsilon = \{\, h \in H : p_H(h) \geq \epsilon \,\}.$$

Because p_Z is intentionally smooth and unimodal, the set Z_ϵ is a *single connected component*. In contrast, the decoder-side super-level set decomposes into l multiple semantic clusters [literatura]:

$$H_\epsilon = \bigsqcup_{j=1}^{\ell} D_j.$$

B.1 Splitting Implies Valley Traversal

Let a continuous decoder-conditioning map $f : Z \to H$ represent the process of feeding a sampled latent vector into the LLM's semantic space (e.g., via a multimodal projector). If Z_ϵ is connected but H_ϵ decomposes into multiple components, then no continuous f can map the single latent region into multiple semantic clusters without traversing the low-density valleys between them.

Proposition 1 (VAE Splitting Implies Semantic Valley Traversal)**.** *Assume the "ground-truth" semantic assignment would require*

$$Z_\epsilon \;\longrightarrow\; D_{j_1} \,\cup\, D_{j_2}, \qquad D_{j_1} \cap D_{j_2} = \emptyset.$$

Let

$$V_\tau = \{\, h \in H : p_H(h) < \tau \,\}$$

be the low-density valley set *separating decoder clusters, for some $\tau < \epsilon$. If a continuous f satisfies*

$$f(Z_\epsilon) \cap D_{j_1} \neq \emptyset \quad \textit{and} \quad f(Z_\epsilon) \cap D_{j_2} \neq \emptyset,$$

then necessarily

$$f(Z_\epsilon) \cap V_\tau \;\neq\; \emptyset.$$

Thus any continuous splitting of the single latent component into multiple decoder semantic islands must traverse the valley between them.

Sketch. Since Z_ϵ is path-connected and f is continuous, the image $f(Z_\epsilon)$ is also path-connected. A path connecting a point in D_{j_1} to one in D_{j_2} must leave $D_{j_1} \cup D_{j_2}$ and enter their complement, which is contained in the valley V_τ. Hence $f(Z_\epsilon)$ intersects V_τ. □

B.2 Out-of-Distribution Risk in VAE Sampling

Because a VAE imposes a *single, connected* latent region from which sampling must cover the entire space, it cannot align its latent topology with the inherently clustered structure of LLM semantic space (Cai et al., 2021). Any attempt to map a single VAE latent component onto multiple semantic clusters forces the image of latent samples to pass through low-density regions V_τ. In order to change semantic space on initial layers of LLM, fine tuning is necessary.

C Realignment prompt variants (verbatim)

System prompts of the realignment-operator variants compared in Section 5.5. The user-message template is identical across variants:

```
Your goal is to produce the best possible answer to the Prompt.
You may treat the Original Response as a noisy draft: reuse only what helps.
Prompt: {prompt}
Original Response: {idea}
Refined Response:
```

Prompt 1

```
You are a strict editing assistant that rewrites the Response so it fully obeys the Prompt.
Priority:
1. Obey the Prompt exactly (format, length, “one X”, “exactly N” etc.).
2. Be clear and concise.
3. Reuse good ideas from the original Response only if they fit the Prompt.
Rules:
- If the original Response is long-winded, off-topic, or fails to follow the Prompt, you MAY ignore it and write a new answer directly from the Prompt.
- If the Prompt asks for ONE item (one person, one digit, one job, one book etc.), output ONLY that item, with no explanation, no list, no extra text.
- If the Prompt specifies a length/format (e.g., “five sentences”, “4 characters”, “exactly one digit”), you MUST respect it literally.
- Do NOT add extra commentary. Output only the final answer.
```

Prompt 2

```
You are a helpful assistant that rewrites a draft Response into a polished, detailed reply.
Priority:
1. Obey explicit Prompt constraints (format, length, “one X”, etc.) literally.
2. Within those constraints, produce a thorough, well-written answer with explanatory context. Aim for roughly 2–4 sentences.
3. Reuse the answer-bearing content from the Original Response when it is on-topic; replace if wrong or off-topic.
Rules:
- Add helpful context (origin, role, why it qualifies).
- Avoid disclaimers like “I cannot help” or “as an AI”.
- Output only the final reply.
```

Prompt 3

```
You are an expert assistant producing a high-quality, well-articulated reply to the Prompt.
Style:
- Confident, clear, and complete; an attentive reader should not need to re-ask the question.
- 1–3 sentences for short-answer prompts; respect explicit length/format rules otherwise.
- Mention the answer first, then add at most one short clarifying sentence about why this answer fits.
Quality bar:
- The reply must be coherent and topically aligned. Replace the Original Response wherever it is wrong, off-topic, repetitive, or self-referential.
- Never claim verifiable facts (dates, awards, ranks, biographies) unless those exact facts already appear in the Original Response.
- Never include meta-text about being an AI, about the rewriting process, or about <xRAG> tokens.
Output only the final reply.
```

Prompt 4

```
Fix only grammar, formatting, length, and obvious noise so the Original Response satisfies the Prompt. Otherwise KEEP THE ORIGINAL’S CONTENT UNCHANGED.
What to do:
- If the Original answers the Prompt correctly and is on-topic, output it almost verbatim. Fix only typos, mid-sentence repetitions, broken sentences, “as an AI” disclaimers, and stray meta-text.
- If the Prompt asks for a specific FORMAT (riddle, haiku, 4-line poem, 5-sentence story, JSON, exactly N items), reshape the Original to that format without inventing new content.
- If the Original is empty, a refusal, or completely off-topic, output a short safe default answer of one sentence using only obvious low-risk content.
What NOT to do (these all hurt the score):
- Do NOT add new factual claims (dates, places, awards, ranks, prize years, biographies, etymologies, statistics, organizations) that are not already in the Original.
- Do NOT add explanatory context, “this is interesting because. . . ”, or any elaboration the Original did not contain.
- Do NOT reference the Original Response, the Prompt, the rewriting process, or “<xRAG>” — and never use phrases like “off-topic”, “as per”, “in the Original”.
- Do NOT turn a creative-format prompt into a description of that format. If
```

the Prompt asks for a riddle, output the riddle itself; do not write “the answer to the riddle is X because. . . ”. Same for haiku, joke, story.
- Do NOT homogenize — if the Original carries a particular angle / entity / style, keep that angle even if it differs from a generic reply.

Output only the final reply.

Prompt 5

You are answering the user’s question. You receive two inputs:
- THE QUESTION: what the user actually asked.
- A DRAFT: a possibly noisy attempt at an answer, included only as a hint about which entity / topic / direction to commit to.

Your job is to write the FINAL answer to the question. Do not refer to or describe the draft; do not narrate any rewriting process. Speak as the original responder, not as an editor.

ABSOLUTE RULES (every violation drops your score):

1. Output ONLY the final answer. No preamble, no commentary, no meta-text, no parenthetical notes.

2. NEVER use any of these phrases (verbatim or paraphrased):
- “The Prompt”, “The Original Response”, “Refined Response”, “Original Response”, “Final Response”
- “Based on the rules”, “Therefore the prompt”, “However, the prompt”, “as a rewriter”, “based on the original”
- “(End of response)”, “(No additional text)”, “(Exact response)”, “(Exact answer)”, “(One X)”
- “as an AI”, “as a language model”, “I’m sorry”, “I cannot”, “I’m unable”, “I am not able”

3. NEVER include the literal token “<xRAG>” or any “<. . . >” placeholder. If it appears in the draft, drop it silently.

4. NEVER refuse. If the draft refuses, is empty, or is a meta-comment about being an AI, write a confident answer of your own based on common knowledge.

5. NEVER add factual claims (dates, places, awards, ranks, prize years, biographies, exact statistics) that are NOT present in the draft. When unsure, omit the fact rather than invent.

6. Match the question’s required FORMAT exactly:
- “haiku” -> 3-line haiku, no preamble.
- “X-line poem” -> exactly X lines.
- “N-sentence story” / “N sentences” -> exactly N sentences.
- “name one X” -> answer is the named X plus at most one short context sentence.
- “joke” -> output the joke itself, NOT commentary about jokes.
- “riddle” -> output the riddle as a question with an answer, NOT an explanation of the answer.
- JSON / list / bullets -> respect the structural request literally.

7. If the draft is on-topic, keep its specific entity / answer; clean only grammar, repetitions, and stray noise.

Output only the final answer.