

CANVAS: ADJUSTABLE CREATIVITY CONTROL IN LARGE LANGUAGE MODELS VIA SPARSE NEURON SCALING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models often produce homogeneous responses to open-ended creative requests, yet increasing variation alone does not ensure meaningful novelty. We study adjustable creativity control without retraining, addressing two challenges: distinguishing novelty from uncontrolled deviation and identifying neurons whose amplification improves creative generation. We introduce CANDLE, an evaluation framework that measures semantic and lexical deviation from the unmodified model's fixed, prompt-specific greedy response under explicit quality constraints. Building on this framework, CANVAS ranks feed-forward neurons by the estimated first-order effect of scaling on a creative-token objective, filters potentially destabilizing targets, and scales a sparse subset of weights. Intervention strength is continuously adjustable, and fixed-strength edits require no additional inference-time computation. Experiments across three model families and three creative-task categories show improved empirical novelty–quality trade-offs. CANDLE's novelty metric achieves agreement scores of 0.760 with LLM judges and 0.669 with human annotators. At a scaling strength of $\alpha = 1.8$, CANVAS increases novelty scores by 14.13% on Qwen and 19.44% on Gemma, with no observed decline in mean quality. Ablations support the contributions of intervention-aligned attribution and stability safeguards. These results establish sparse neuron scaling as a practical approach to adjustable, quality-constrained creative generation.

https://anonymous.4open.science/r/CANVAS-Steer/

1 INTRODUCTION

Open-ended creative requests, including story writing, brainstorming, and role-play, are common in real-world interactions with LLM assistants (Zheng et al., 2024; Zhao et al., 2024; Jiang et al., 2026). These tasks require outputs that are both *novel* and *appropriate*: responses should depart from familiar patterns while respecting the user's intent and task constraints (Runco & Jaeger, 2012; Boden, 2004). Yet LLM outputs often exhibit homogenization (Jiang et al., 2026), and AI assistance can reduce the diversity of human writing (Padmakumar & He, 2024; Doshi & Hauser, 2024). Moreover, the desired degree of novelty varies across tasks and users: exploratory brainstorming may favor unconventional ideas, whereas tightly constrained writing prioritizes adherence to specific requirements. Although post-training can change creative tendencies (Mohammadi, 2024), it requires additional training cost (Hu et al., 2022) and lacks the flexibility to dynamically adjust the degree of creativity. We therefore seek *adjustable creativity control without retraining* to flexibly modulate meaningful novelty while preserving generation quality.

Interventions on model internals offer a route to controlling model behavior at different levels, from hidden representations to individual neurons, attention heads, and multi-component circuits (Turner et al., 2024b; Dai et al., 2022; Wang et al., 2023). We focus on feed-forward network (FFN) neurons, motivated by evidence that targeted interventions can alter knowledge expression and generation behavior (Dai et al., 2022; Geva et al., 2021). Building on these findings, we investigate whether scaling a sparse subset of FFN neurons can reliably increase novelty while preserving generation quality. This formulation enables adjustable intervention strength, with fixed-strength scaling incorporated into model weights without additional inference-time computation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

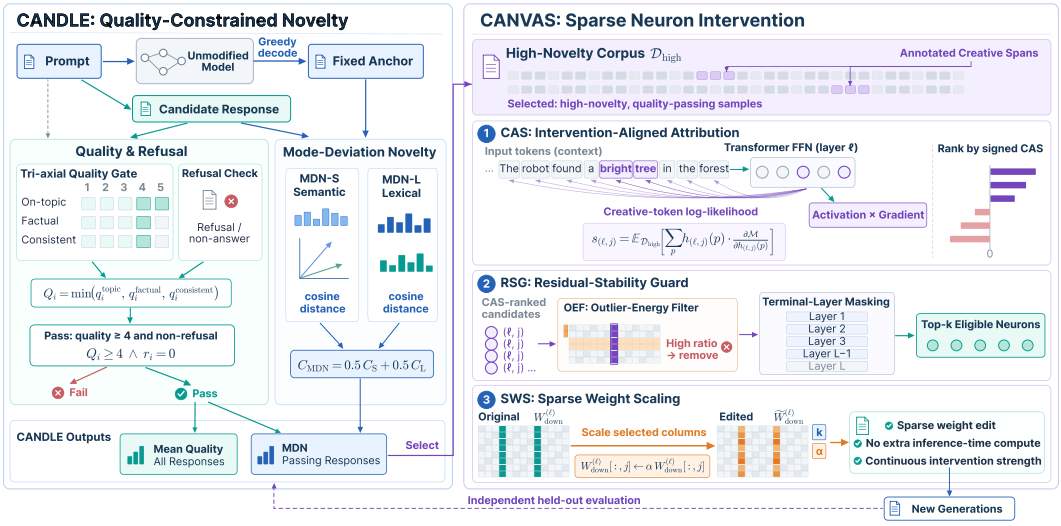


Figure 1: Overview of CANDLE and CANVAS. **Left:** CANDLE measures semantic and lexical novelty relative to the unmodified model's fixed greedy response, reporting novelty over quality-passing, non-refusal responses alongside mean quality over all responses. **Right:** CANVAS uses annotated creative spans from a high-novelty corpus to rank FFN neurons by intervention-aligned first-order attribution, filters candidates through outlier-energy filtering and terminal-layer masking, and scales selected weight columns. The number of selected neurons k and the scaling factor α provide continuous intervention-strength control without retraining or inference-time computation.

However, two challenges make such control difficult. **First, existing evaluations provide limited signals for identifying creative gains.** Current metrics struggle to reliably distinguish useful novelty from uncontrolled deviation. Human assessment is grounded but costly (Amabile, 1982), LLM judges suffer from inherent biases (Zheng et al., 2023; Wang et al., 2024), and automatic metrics (*e.g.*, Self-BLEU) capture surface diversity without ensuring task appropriateness (Zhu et al., 2018; Beaty & Johnson, 2021). Without explicit novelty references and quality constraints, these evaluations cannot reliably guide localization or assess interventions. **Second, existing localization methods often fail to identify effective intervention targets.** Methods such as H-Neurons (Gao et al., 2025) and CreativityNeuro (Schapiro et al., 2026) identify *associated* neurons, yet association does not guarantee that amplification will improve generation. Additionally, aggregating activations over long responses obscures fine-grained signals, and modifying polysemantic neurons (Huang et al., 2024) risks disrupting other behaviors. Effective control therefore requires intervention-aware localization and safeguards against unintended degradation.

We address these challenges with a unified evaluation and intervention framework (Figure 1). To evaluate generation, we propose **CANDLE** (Creativity via Anchored Novelty and Degeneracy-guarded Long-form Evaluation) that anchors comparisons to the unmodified model's fixed, prompt-specific greedy response. It quantifies semantic and lexical divergence via **Mode-Deviation Novelty (MDN)**, while a strict quality gate filters irrelevant or hallucinated outputs. CANDLE thus rewards quality-constrained novelty rather than uncontrolled deviation. For intervention, we propose **CANVAS** (Causal Attribution of Neurons and Variance-guarded Amplification by Scaling), which localizes FFN neurons using a first-order approximation of scaling effects on a creative-token proxy, overcoming the limits of simple correlation methods. It applies outlier filtering and terminal-layer masking to prevent degradation, while Sparse Weight Scaling amplifies selected neurons via a continuous control parameter. This enables dynamic creativity adjustment without retraining; moreover, the scaled weights merge directly into the model, eliminating inference-time overhead. We use CANDLE to verify that CANVAS significantly improves the novelty–quality frontier.

We evaluate our framework on three model families across three categories of creative generation. Validation on blinded pairs yields MDN–LLM and MDN–human consistency scores of 0.760 and 0.669, respectively, supporting CANDLE's alignment with both automated and human judgments. CANVAS improves the empirical novelty–quality frontier, achieving MDN gains of 14.13% on Qwen and 19.44% on Gemma at a neuron-scaling strength of $\alpha = 1.8$, with no observed decline

in mean quality. Across six LLM judges, its strongest evaluated configuration ranks highest in comparisons against the greedy anchor, while ablations demonstrate the contributions of targeted attribution and stability safeguards. Mechanistic analyses reveal model-specific neuron distributions, front-loaded attribution during generation, and reduced attribution to selected neurons following safety SFT. Together, these findings support sparse internal intervention as a practical approach to adjustable, quality-constrained creative generation.

2 Related Work

Creativity in LLMs. LLM outputs exhibit homogenization, and AI assistance can reduce collective writing diversity (Padmakumar & He, 2024; Doshi & Hauser, 2024; Jiang et al., 2026). Creativity requires both novelty and appropriateness (Boden, 2004); existing evaluations include divergent-thinking tasks and judged creative writing (Olson et al., 2021; Bellemare-Pepin et al., 2026; Paech, 2024). Automatic measures such as distinct-n, Self-BLEU, MAUVE, and embedding diversity quantify variation or distributional similarity (Li et al., 2016; Zhu et al., 2018; Pillutla et al., 2021; Jiang et al., 2026), but do not jointly measure prompt-specific, default-relative novelty and task appropriateness. LLM judges offer sample-level assessments but introduce cost, variability, and biases (Zheng et al., 2023; Liu et al., 2023; Wang et al., 2024; Bavaresco et al., 2025). CANDLE anchors semantic and lexical novelty to a fixed greedy response and applies a task-quality gate, supporting controlled comparisons across interventions within each model and prompt setting. Its MDN score measures default-relative novelty, not creativity in its entirety.

Behavioral Control and Neuron-Level Intervention. Decoding methods adjust the quality–diversity balance through sampling (Fan et al., 2018; Holtzman et al., 2020; Meister et al., 2023; Hewitt et al., 2022; Zhang et al., 2021), while activation steering modifies internal representations at inference time (Li et al., 2023; Zou et al., 2025; Turner et al., 2024a; Rimsky et al., 2024). Weight-space methods make behavioral or knowledge changes (Ilharco et al., 2023; Meng et al., 2022). CANVAS uses sparse FFN weight scaling for continuous strength control, requiring no additional inference-time computation. Neuron localization draws on evidence linking FFN neurons to knowledge, skills, and behavior (Geva et al., 2021; Dai et al., 2022; Wang et al., 2022; Gao et al., 2025). However, magnitude-only scores omit intervention direction, and activation pooling obscures token-localized signals (Hewitt & Liang, 2019; Dobrzeniecka et al., 2025). CAS uses gradient–activation products to estimate scaling's first-order effect on a creative-token proxy, preserving token resolution and sign; stability safeguards restrict potentially destabilizing modifications (Sun et al., 2024; Zhang & Sennrich, 2019). Concurrent work also scales creativity-associated FFN neurons, using Wanda-style importance, with evaluation on short cognitive instruments (Schapiro et al., 2026). CANVAS instead uses signed attribution and explicit residual-stability safeguards.

3 Creativity Quantification: The CANDLE Protocol

Measuring creativity requires separating novelty from uncontrolled variation. We introduce **CANDLE**, which measures *default-relative novelty under quality constraints* by combining semantic and lexical deviation from the unmodified model's greedy response with a separate quality gate.

3.1 Data and Sampling Protocol

Prompt construction. We draw creativity-related prompts from the Infinite-Chats Taxonomy dataset (Jiang et al., 2026) across three buckets: *speculative* (hypothetical scenarios), *ideation* (brainstorming and listing), and *pure writing* (stories, lyrics, and role-play). Localization uses 500 prompts per bucket with ten generations each, totaling $1{,}500$ prompts and $15{,}000$ responses. Appendix A.1 details filtering, stratification, and token annotation. Test prompts come from the remaining pool, excluding prior generation or training prompts and near-duplicates. We retain only candidates whose greedy baseline answers without refusal and passes the quality gate, yielding 50 prompts per bucket (150 total). Evaluation thus targets baseline-answerable prompts (Appendix A.2).

Generation and anchoring. We use temperature 1.0, top-p 0.9, top-k 50, and an $8{,}192$-token generation limit, sampling 5 responses per test prompt. Settings are shared except for decoding-baseline parameters; prompts and seeds are matched across methods (Appendix A.3). For each model and

prompt p, the unmodified model generates a single greedy anchor b_p, fixed across interventions to provide a common reference for behavioral shifts.

3.2 METRIC DEFINITION

CANDLE combines refusal identification, a Tri-axial Quality Gate, and Mode-Deviation Novelty (MDN). Refusal detection uses a rule-based matcher (Appendix A.5).

Tri-axial Quality Gate. Compared with open-ended creativity assessment, quality evaluation can rely on more explicit criteria. Prior work shows that LLM judges guided by structured rubrics can achieve improved agreement with human quality judgments (Liu et al., 2023). We therefore use GPT-4o, with temperature 0 and structured JSON output, to score three dimensions from 1 to 5: `on_topic` (fulfillment of prompt requirements), `factual` (correctness of claims where invention is not invited), and `consistent` (coherence and absence of contradictions, truncation, or degeneration). Fictional and hypothetical content is assessed against the task's premises. The judge is instructed to evaluate appropriateness, not originality or verbosity; the rubric appears in Appendix A.6. Bottleneck quality prevents one dimension from compensating for failure in another:

$$Q_i = \min\left\{q_i^{\text{topic}}, q_i^{\text{factual}}, q_i^{\text{consistent}}\right\}, \qquad \overline{Q} = \frac{1}{N}\sum_{i=1}^{N} Q_i. \tag{1}$$

Reported **mean quality** includes all responses, including refusals and gate failures. With refusal indicator r_i, the gate and its pass rate are

$$g_i = \mathbb{1}[Q_i \geq 4]\mathbb{1}[r_i = 0], \qquad \widehat{\pi}_{\text{gate}} = \frac{1}{N}\sum_{i=1}^{N} g_i. \tag{2}$$

MDN: Mode-Deviation Novelty. For response y_i to prompt $p(i)$, we measure default-relative novelty along two dimensions: semantic deviation (**MDN-S**), using ℓ_2-normalized `text-embedding-3-small` representations $e(\cdot)$ (OpenAI, 2024), and lexical deviation (**MDN-L**), using ℓ_2-normalized unigram and bigram TF-IDF vectors $t(\cdot)$:

$$C_{\text{S}}(i) = \max\{0, 1 - e(y_i)^\top e(b_{p(i)})\}, \quad C_{\text{L}}(i) = \max\{0, 1 - t(y_i)^\top t(b_{p(i)})\}. \tag{3}$$

Here, $b_{p(i)}$ is the unmodified model's greedy response to prompt $p(i)$, generated once and fixed across interventions as a shared novelty anchor. The TF-IDF vocabulary and IDF weights are likewise fixed across interventions, and responses and anchors follow identical preprocessing (Appendix A.7). We combine the components without performance-based tuning:

$$C_{\text{MDN}}(i) = \lambda C_{\text{S}}(i) + (1-\lambda) C_{\text{L}}(i), \qquad \lambda = 0.5. \tag{4}$$

In our main results, MDN is averaged over responses that pass the quality gate:

$$C_{\text{MDN}} = \frac{\sum_i g_i C_{\text{MDN}}(i)}{\sum_i g_i}. \tag{5}$$

Higher MDN scores indicate greater default-relative novelty, capturing the novelty component of creativity. We report this score alongside mean quality over all responses. Because gate filtering changes the evaluated response set and gives greater weight to prompts with more passing responses, we interpret conditional novelty gains jointly with quality scores and gate-pass diagnostics. A comparison of CANDLE with alternative creativity metrics is provided in Appendix A.8.

3.3 METRIC VALIDATION

Pairwise evaluation. We assess creativity using pairwise preferences rather than absolute ratings, following established LLM-as-judge protocols (Zheng et al., 2023). To improve evaluation stability, LLM judges assess each pair in both presentation orders, with method identities hidden. Preferences for the target response are encoded as 1 (win), 0.5 (tie), or 0 (loss), then averaged across orders to mitigate position bias (Wang et al., 2024). The evaluation prompt appears in Appendix A.4.

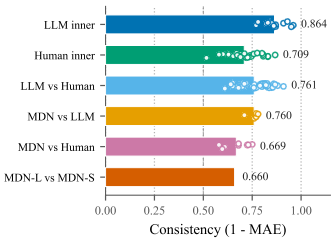


Figure 2: Pairwise consistency on comparisons. Bars show means; dots show judge pairs.

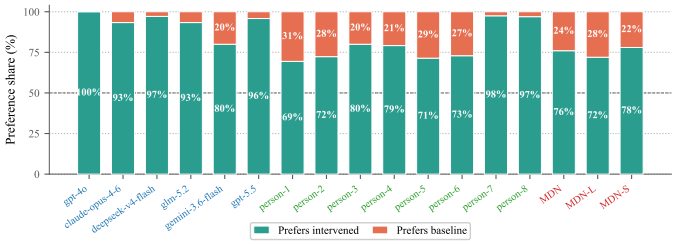


Figure 3: Preference shares on the same comparisons, excluding ties: outputs from the model modified by our method (green) versus the unmodified model (orange).

Convergent validity. Six LLM judges and eight human annotators evaluate 50 blinded pairs, each comprising an intervened output and a sampled output from the unmodified model. The human evaluation protocol is detailed in Appendix C.3. We quantify preference-score agreement using one minus the mean absolute error (MAE):

$$A(a,b) = 1 - \frac{1}{N}\sum_{i=1}^{N} |s_{a,i} - s_{b,i}|, \qquad N = 50. \tag{6}$$

This measure assigns full agreement to identical judgments, partial agreement to tie–win or tie–loss differences, and no agreement to opposing preferences.

Figure 2 shows that MDN exhibits substantial consistency with both LLM and human judgments. MDN–LLM consistency (0.760) is comparable to LLM–human consistency (0.761), while MDN–human consistency (0.669) approaches the inter-human reference (0.709). The lower consistency between MDN-S and MDN-L (0.660) suggests that the two components capture partly distinct aspects of response variation, consistent with their semantic and lexical formulations. Figure 3 further shows that MDN shares the directional preference of LLM judges and human evaluators for outputs from our modified model over sampled outputs from the unmodified model. Excluding ties, LLM and human preference shares are 80–100% and 69–98%, respectively, compared with 76% for MDN, 72% for MDN-L, and 78% for MDN-S. Together, these results support convergence in both pairwise judgments and overall preference direction.

We further assess resistance to metric gaming and sensitivity to the semantic–lexical mixing weight. CANDLE rejects 100% off-topic and degenerate adversarial responses and 80% of paraphrase-only responses, outperforming the tested baselines (Appendix B.1). Rankings remain stable under moderate changes in λ, supporting $\lambda = 0.5$ as a fixed, symmetric design choice (Appendix B.2).

4 Locating Creativity Neurons: CANVAS

CANVAS converts creativity-related supervision into a sparse weight intervention through three stages (Figure 1, right): **Creativity Attribution Score** (CAS) estimates the local effect of neuron amplification on a differentiable proxy; **Residual-Stability Guard** (RSG) screens potentially destabilizing candidates; and **Sparse Weight Scaling** (SWS) applies the edit. The guiding principle is to *localize creativity-related neurons using the derivative of the intervention to be deployed.*

4.1 CAS: Intervention-Aligned First-Order Attribution

Effective neuron-level creativity control requires localization that preserves token-level signals and predicts the effect of amplification. Existing approaches face two limitations. First, association-based scores do not establish whether amplification would promote or suppress the target behavior, and magnitude-only scores provide no directional information (Hewitt & Liang, 2019; Dobrzeniecka et al., 2025). Second, conventional activation-based localization often pools neuron activity across tokens before measuring its association with a behavior (Dai et al., 2022; Wu et al., 2024; Gurnee et al., 2023), potentially obscuring localized contributions in long-form generation.

To align localization with the intended intervention, CAS differentiates a creativity-related objective with respect to neuron scaling. From the generations in §3.1, we construct a quality-gated high-

novelty subset $\mathcal{D}_{\text{high}}$ with annotated creative spans (Appendix A.1). These spans define a log-likelihood proxy conditioned on the observed token prefixes, from which CAS computes signed, position-specific contributions before aggregation. The resulting score estimates the local effect of neuron amplification on this proxy, rather than merely its association with high-novelty responses.

Creative-token objective. For a prompt x and a high-novelty response r, let $\mathcal{T}_r$ denote the token positions within annotated creative spans. We define a log-likelihood proxy conditioned on the observed token prefixes:

$$\mathcal{M}(x, r) = \frac{1}{|\mathcal{T}_r|} \sum_{t \in \mathcal{T}_r} \log p_\theta(r_t \mid x, r_{<t}). \tag{7}$$

This objective measures the mean log-likelihood of annotated creative tokens given the prompt and their preceding tokens.

Derivative of neuron scaling. To identify neurons whose amplification would promote annotated creative tokens, we estimate how the creative-token objective changes under the intended scaling intervention. Let $h_{\ell j}(p)$ denote neuron j's activation at position p immediately before the down projection in layer ℓ (after the gating product in gated FFNs). Its contribution to the FFN output is $h_{\ell j}(p) W^{(\ell)}_{\text{down}}[:, j]$. Multiplying this contribution by a scalar $\alpha_{\ell j}$ at every position yields

$$\left.\frac{\partial \mathcal{M}}{\partial \alpha_{\ell j}}\right|_{\boldsymbol{\alpha}=\mathbf{1}} = \sum_p h_{\ell j}(p) \frac{\partial \mathcal{M}}{\partial h_{\ell j}(p)}. \tag{8}$$

The sum includes all positions that influence the objective, including prompt and preceding response positions, not just the annotated target positions $\mathcal{T}_r$. CAS averages this derivative over the high-novelty corpus:

$$s_{\ell j} = \mathbb{E}_{(x,r)\sim\mathcal{D}_{\text{high}}} \left[\sum_p h_{\ell j}(p) \frac{\partial \mathcal{M}(x, r)}{\partial h_{\ell j}(p)} \right]. \tag{9}$$

For a fixed neuron set $\mathcal{N}$ scaled by a common factor α, define the corpus-averaged objective as $\overline{\mathcal{M}}(\alpha) = \mathbb{E}_{(x,r)\sim\mathcal{D}_{\text{high}}}[\mathcal{M}_\alpha(x, r)]$, where $\mathcal{M}_\alpha$ is evaluated under the scaled model. Its first-order expansion is

$$\overline{\mathcal{M}}(\alpha) = \overline{\mathcal{M}}(1) + (\alpha - 1) \sum_{(\ell,j)\in\mathcal{N}} s_{\ell j} + O\big((\alpha - 1)^2\big). \tag{10}$$

Thus, CAS connects neuron selection directly to the predicted local effect of the deployed intervention. Positive scores predict that amplification increases the proxy, and negative scores predict a decrease. We therefore rank neurons by signed CAS rather than its magnitude. Although gradient–activation products are established in attribution and pruning (Molchanov et al., 2019; Dalvi et al., 2019), here they arise directly from differentiating the deployed scaling operator. The derivative identity is exact, whereas the first-order approximation is valid only locally.

4.2 RSG: Residual-Stability Guard

A favorable local derivative does not guarantee quality preservation under finite amplification. RSG therefore applies two empirical safeguards to restrict potentially destabilizing edits.

Outlier-energy filtering. LLMs can exhibit unusually large activations along certain residual dimensions (Sun et al., 2024). Perturbations along these dimensions can change the RMSNorm normalization factor, thereby affecting other coordinates (Zhang & Sennrich, 2019). We therefore screen neurons whose output directions concentrate on high-RMS residual dimensions.

Let od_ℓ denote the top-m residual dimensions ranked by RMS activation, estimated on calibration data using the unmodified model. The **Outlier-Energy Filter** (OEF) computes

$$\text{OE}(\ell, j) = \frac{\left\| W^{(\ell)}_{\text{down}}[od_\ell, j] \right\|_2^2}{\left\| W^{(\ell)}_{\text{down}}[:, j] \right\|_2^2} \tag{11}$$

and excludes candidates with $\text{OE}(\ell, j) > \tau$. OEF filters out neurons whose output directions concentrate excessively on outlier dimensions.

Terminal-layer masking. Prior work characterizes transformer layers as progressively refining next-token predictions (Belrose et al., 2025). Late-layer perturbations undergo fewer subsequent transformations before reaching the output distribution, motivating caution when amplifying neurons in these layers. We therefore apply **Terminal-Layer Masking** to exclude the final L_{mask} FFN layers from scaling and limit potential disruption to output generation.

4.3 SWS: Sparse Weight Scaling

We first select the top k neurons ranked by signed CAS, then apply RSG to obtain the final intervention set $\mathcal{N}_k$. SWS applies the scaling operator differentiated by CAS:

$$W_{\text{down}}^{(\ell)}[:, j] \leftarrow \alpha W_{\text{down}}^{(\ell)}[:, j], \qquad (\ell, j) \in \mathcal{N}_k. \tag{12}$$

Scaling a column multiplies the corresponding neuron's FFN contribution at each position: $\alpha = 1$ leaves it unchanged, and $\alpha > 1$ amplifies it. With the localization procedure and guard settings fixed, SWS offers two controls—the number of selected neurons k and the scaling factor α—with three practical properties: *No additional inference-time computation.* The edit preserves the architecture and tensor shapes, yielding a standard checkpoint without activation hooks or additional decoding operations. *Sparse, activation-dependent modification.* Only $|\mathcal{N}_k|$ output columns are modified, and each neuron's immediate contribution change scales with its activation. *Continuous intervention strength.* Varying α continuously adjusts intervention strength, enabling exploration of the novelty–quality trade-off.

Together, CAS ranks neurons by their predicted local response to scaling, RSG screens potentially destabilizing targets, and SWS converts the selected interventions into deployable checkpoints. Independent generation evaluation measures the resulting novelty gains and quality trade-offs.

5 Experiments

We use the data construction and generation settings in §3.1. LLM evaluation follows the blinded, order-balanced pairwise protocol in §3.3, comparing each response with its fixed greedy anchor.

Models. We evaluate three instruction-tuned open-weight models: `Qwen3-8B`, `Llama-3.1-8B-Instruct`, and `Gemma-3-12B-IT`. These models span different families and architectures. Architectural specifications appear in Appendix C.1.

Baselines. We compare **CANVAS** against the unmodified model (**Base**) and interventions at three levels: *input-level* **Prompting** with explicit creativity instructions; *decoding-level* **Temperature** sampling and **Top-H** decoding; and *activation-level* methods, including **CAA** steering (Rimsky et al., 2024), **Random** (scaling randomly selected neurons), **H-Neurons** (localization using CETT scores and sparse logistic regression) (Gao et al., 2025), and **CreativityNeuro** (localization using Wanda-style parameter importance scores) (Schapiro et al., 2026). For H-Neurons and CreativityNeuro, we implement only their neuron-localization procedures and use our dataset rather than the baselines' original datasets, enabling controlled comparisons of localization strategies.

5.1 CANVAS Performance

Creativity–quality trade-offs. As shown in Figure 4, CANVAS exhibits a monotonic increase in MDN with amplification strength across the tested range on all three models, supporting controllable creativity modulation. On Qwen and Gemma, it offers the strongest creativity gains among evaluated methods at approximately preserved mean quality: at $\alpha = 1.8$, MDN increases by 14.13% and 19.44%, respectively, without an observed decline in mean quality. Among the evaluated neuron-localization methods, CANVAS achieves state-of-the-art creativity–quality trade-offs across all three models. Random scaling also increases creativity in some settings, indicating that generic weight perturbations can shift default behavior; however, the stronger trade-offs achieved by CANVAS support the value of targeted neuron selection over indiscriminate amplification.

Cross-judge preferences. Figure 5 shows variation in absolute scores across judges, with DeepSeek generally scoring lower and GPT-5.5 higher. Nevertheless, CANVAS achieves the highest preference scores under all six LLM judges, with scores ranging from 0.8862 to 0.9727. Across CANVAS

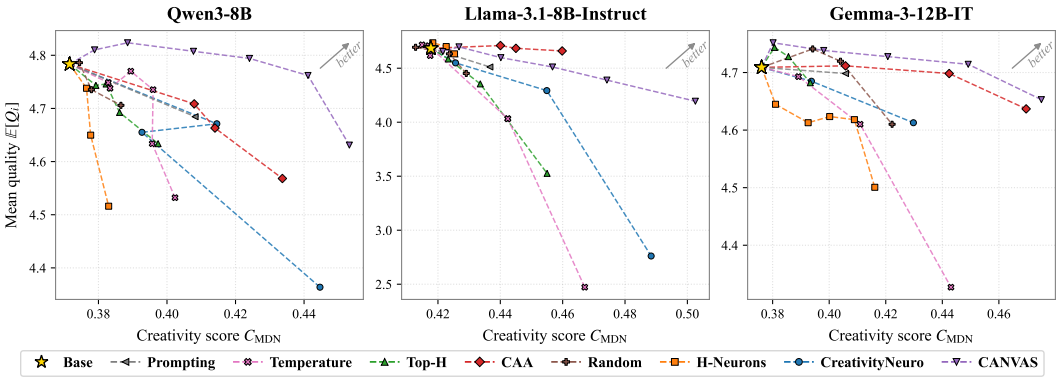


Figure 4: Quality–creativity trade-off across eight steering methods on three models. Each dashed curve traces a method across its scaling range; the gold star marks the unsteered baseline.

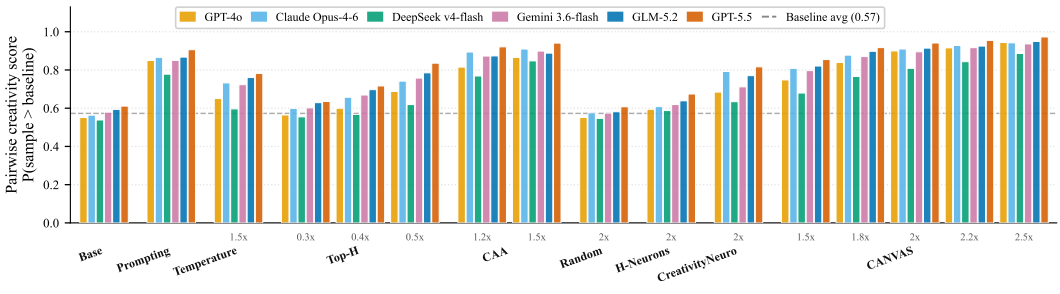


Figure 5: Creativity preference scores against the fixed greedy anchor, evaluated by six LLM judges.

scaling strengths, all LLM scores increase, indicating a shared preference trend toward amplified outputs over the fixed greedy anchor. This cross-evaluator convergence supports perceived creativity gains beyond an increase in the automatic novelty score.

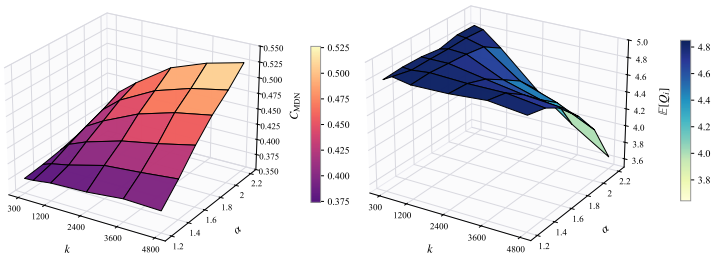
Hyperparameter analysis. Figure 7 shows that k and α jointly control the creativity–quality trade-off. Increasing either generally raises MDN, while substantial quality degradation occurs when both are large, reaching 3.64 at $k = 4800$, $\alpha = 2.2$. A practical operating region is $k = 1200$, $\alpha \in [2.0, 2.2]$, which yields substantial novelty gains while keeping mean quality above 4.6. Thus, intervention breadth and strength can be tuned jointly to balance novelty and quality.

Ablation study. Figure 8 shows that the full configuration achieves the most favorable overall creativity–quality trade-off. Alternative OEF settings yield weaker trade-offs, while removing TLM causes greater quality degradation at stronger amplification without a corresponding novelty gain. Replacing annotated creative spans with all tokens or randomly selected tokens substantially limits MDN gains, highlighting the importance of a targeted attribution objective. Together, these results support the complementary roles of creative-token supervision and residual-stability safeguards.

Additional analyses. Additional results include detailed results for Figure 4 (Appendix D.1), evaluator agreement (Appendix D.2), semantic and lexical novelty analyses (Appendix D.3), OEF and TLM ablations (Appendix D.4), out-of-distribution generalization (Appendix D.5), and response distribution visualizations (Appendix D.6). Qualitative case studies further illustrate how the interventions affect generated content (Appendix E).

5.2 Mechanistic Analysis

Layerwise distribution of creativity neurons. Figure 9 reveals model-specific localization patterns. Qwen exhibits a broad concentration in middle-to-late layers with few early-layer neurons, whereas Llama shows a more distributed profile with elevated counts near the final layers. Gemma combines a mid-layer concentration with a pronounced first-layer peak. The substantial middle-to-late-layer presence is consistent with a role for semantic composition and content selection, but the early-layer contributions caution against reducing creativity to exclusively high-level processing.



(a) Creativity score C_{MDN}. (b) Mean_quality $\mathbb{E}[Q_i]$.

Figure 7: Joint effect of the number of intervened neurons k and activation scale factor α on output creativity and quality.

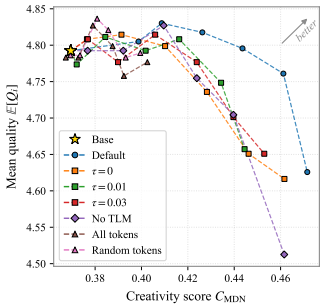


Figure 8: RSG and token-selection ablations.

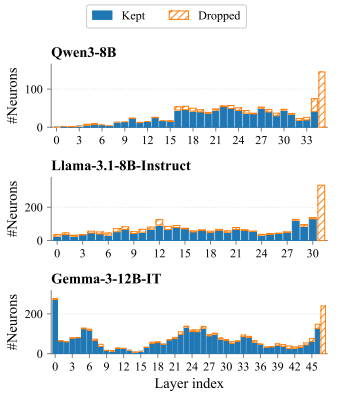


Figure 9: Layer-wise distribution of creativity neurons in three models.

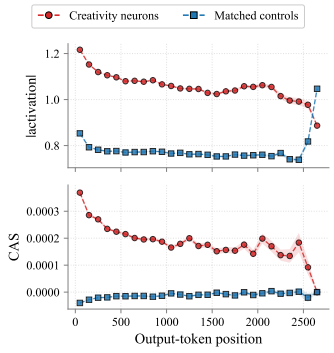


Figure 10: Mean |activation| (top) and CAS (bottom) vs. output-token position on Qwen3-8B.

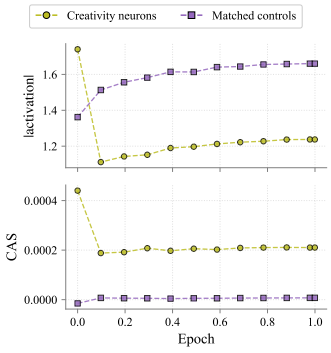


Figure 11: Mean |activation| (top) and CAS (bottom) across Qwen3-8B checkpoints SFT-tuned on WildJailbreak.

Creativity-related influence is front-loaded. Figure 10 compares the creativity neurons with an equal-sized, layer-wise matched control set. Creativity neurons exhibit consistently larger activation magnitudes and more positive CAS over most of the response. The CAS gap is greatest near response onset and narrows as generation proceeds, whereas the activation-magnitude gap persists. This pattern suggests that the selected neurons exert their strongest influence on the creative-token objective early in generation, rather than contributing uniformly throughout the response.

Alignment selectively attenuates creativity-neuron signals. We perform safety SFT on WildJailbreak (Appendix C.5), training the model to refuse harmful requests while responding to benign ones, and track neuron activity across checkpoints. Figure 11 shows a sharp early reduction in both activation magnitude and CAS for the identified creativity neurons, followed by partial recovery and stabilization below their pre-SFT levels. These contrasting trajectories suggest that safety SFT selectively attenuates creativity-related signals rather than uniformly suppressing neuronal activity. Control construction and experiment-specific settings are detailed in Appendix C.4.

6 Conclusion

We introduced CANDLE and CANVAS for evaluating and controlling long-form LLM creativity without retraining. CANDLE measures default-relative semantic and lexical novelty under quality constraints. CANVAS combines intervention-aligned neuron attribution, stability safeguards, and sparse weight scaling to provide adjustable control without additional inference-time computation. Experiments across three model families show improved novelty–quality trade-offs, while human and LLM judgments support CANDLE's validity. Ablations and out-of-distribution evaluations further support the approach. Together, these findings establish sparse neuron scaling as a practical tool for creative generation and underscore the importance of assessing novelty alongside quality.

AI USE STATEMENT

Generative AI systems were used both as objects of study and as research tools. As described in the paper, language models generated the experimental responses, GPT-4o provided structured quality judgments and creative-token annotations, and six language-model judges participated in the pairwise evaluation. Generative AI coding assistants were also used to help implement and debug data-processing, evaluation, and visualization scripts; to suggest revisions to experimental analyses; and to polishing the manuscript. The authors determined the research questions, methodology, and reported claims; inspected the underlying samples and aggregate outputs; executed and tested the analysis code; and reviewed all AI-assisted text and code before inclusion. No AI-generated suggestion was treated as evidence without verification against the experimental records or cited source. The authors take full responsibility for the accuracy, originality, and integrity of the submission and for all artifacts produced with generative-AI assistance.

ETHICS STATEMENT

We have read and adhered to the ICLR Code of Ethics. The human evaluation was limited to a blinded preference task involving AI-generated responses to open-ended creative prompts. Eight non-author annotators provided informed consent, completed qualification and attention checks, and were instructed not to use AI assistance. Method identities and response order were concealed, and the paper reports judgments only in aggregate; further procedural details are given in Appendix C.3. The remaining experiments use publicly available model checkpoints and research datasets, which should be used in accordance with their respective licenses and terms.

This work measures and modifies model behavior and therefore presents potential dual-use risks. Sparse weight interventions that increase creative variation could also amplify misleading, unsafe, stereotyped, or otherwise undesirable content, and the quality gate and safety-SFT analyses in this work do not constitute a comprehensive safety evaluation. We therefore do not recommend deploying edited checkpoints without application-specific safety, bias, and robustness testing. Creativity judgments are also culturally and linguistically situated: both human and model evaluators may encode biases, and default-relative novelty should not be interpreted as an absolute measure of artistic or social value. We mitigate these limitations through blinded comparisons, multiple model and human evaluators, explicit quality criteria, and transparent reporting of the evaluation protocol.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce the study across the main paper, appendices, supplementary material, and our open-source code release. Section 3.1 defines the data splits, sampling parameters, fixed anchors, and matched random seeds; Section 3.2 specifies TQG and MDN; and Section 4 gives the CANVAS objective, neuron-selection procedure, safeguards, and weight intervention. Appendix A documents prompt stratification, sample construction, refusal detection, scoring prompts, and anchor design; Appendix C.1 records the models and numerical settings; and Appendix C.2 provides baseline implementations and parameter grids. The LLM-judge and human-evaluation protocols are reproduced in Appendices A.4 and C.3, and complete parameter-level results appear in Appendix D.1.

We have open-sourced all code at https://anonymous.4open.science/r/CANVAS-Steer/, including scripts for data construction, neuron attribution and selection, model intervention, baseline implementation, generation, evaluation, statistical analysis, ablations, and figure production. We also provide intermediate results and cached evaluator outputs where redistribution is permitted. These artifacts retain sample identifiers and configuration metadata, enabling independent reconstruction of matched comparison sets, quality gating, MDN aggregation, ablations, and all reported figures. We additionally document model- and API-dependent components explicitly because hosted evaluators may exhibit residual nondeterminism even under fixed decoding settings.

REFERENCES

Teresa M Amabile. Social psychology of creativity: A consensual assessment technique. *Journal of personality and social psychology*, 43(5):997, 1982.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 238–255, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-252-7. doi: 10.18653/v1/2025.acl-short.20. URL https://aclanthology.org/2025.acl-short.20/.

Roger E Beaty and Dan R Johnson. Automating creativity assessment with semdis: An open platform for computing semantic distance. *Behavior research methods*, 53(2):757–780, 2021.

Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A Olson, Yoshua Bengio, and Karim Jerbi. Divergent creativity in humans and large language models. *Scientific Reports*, 16(1):1279, 2026.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens, 2025. URL https://arxiv.org/abs/2303.08112.

Margaret A Boden. *The creative mind: Myths and mechanisms*. Routledge, 2004.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642731. URL https://doi.org/10.1145/3613904.3642731.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.581. URL https://aclanthology.org/2022.acl-long.581/.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.33016309. URL https://doi.org/10.1609/aaai.v33i01.33016309.

Alicja Dobrzeniecka, Antske Fokkens, and Pia Sommerauer. Improving causal interventions in amnesic probing with mean projection or LEACE. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 12981–12993, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.674. URL https://aclanthology.org/2025.findings-acl.674/.

Anil R. Doshi and Oliver P. Hauser. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290, 2024. doi: 10.1126/sciadv.adn5290. URL https://www.science.org/doi/abs/10.1126/sciadv.adn5290.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082/.

Cheng Gao, Huimin Chen, Chaojun Xiao, Zhiyi Chen, Zhiyuan Liu, and Maosong Sun. H-neurons: On the existence, impact, and origin of hallucination-associated neurons in llms, 2025. URL https://arxiv.org/abs/2512.01797.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing, 2023. URL https://arxiv.org/abs/2305.01610.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

John Hewitt, Christopher Manning, and Percy Liang. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3414–3427, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.249. URL https://aclanthology.org/2022.findings-emnlp.249/.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. RAVEL: Evaluating interpretability methods on disentangling language model representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8669–8687, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.470. URL https://aclanthology.org/2024.acl-long.470/.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and Nouha Dziri. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models, 2024. URL https://arxiv.org/abs/2406.18510.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=saDOrrnNTz.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 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*, pp. 110–119, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1014. URL https://aclanthology.org/N16-1014/.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=aLLuYpn83y.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121, 2023. doi: 10.1162/tacl_a_00536. URL https://aclanthology.org/2023.tacl-1.7/.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=-h6WAS6eE4.

Behnam Mohammadi. Creativity has left the chat: The price of debiasing language models, 2024. URL https://arxiv.org/abs/2406.05587.

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11256–11264, 2019. doi: 10.1109/CVPR.2019.01152.

Jay A. Olson, Johnny Nahas, Denis Chmoulevitch, Simon J. Cropper, and Margaret E. Webb. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*, 118 (25):e2022340118, 2021. doi: 10.1073/pnas.2022340118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2022340118.

OpenAI. New embedding models and API updates, January 2024. URL https://openai.com/index/new-embedding-models-and-api-updates/.

Vishakh Padmakumar and He He. Does writing with language models reduce content diversity? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Feiz5HtCD0.

Samuel J. Paech. Eq-bench: An emotional intelligence benchmark for large language models, 2024. URL https://arxiv.org/abs/2312.06281.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: measuring the gap between neural text and human text using divergence frontiers. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Mark A Runco and Garrett J Jaeger. The standard definition of creativity. *Creativity research journal*, 24(1):92–96, 2012.

Samuel Schapiro, Core Francisco Park, Felix Sosa, and Lav R. Varshney. Creativityneuro: Steering language model weights to improve divergent thinking and reduce mode collapse, 2026. URL https://arxiv.org/abs/2607.01433.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=F7aAhfitX6.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024a. URL https://arxiv.org/abs/2308.10248.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024b. URL https://arxiv.org/abs/2308.10248.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. Finding skill neurons in pre-trained transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11132–11152, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.765. URL https://aclanthology.org/2022.emnlp-main.765/.

Xuansheng Wu, Wenlin Yao, Jianshu Chen, Xiaoman Pan, Xiaoyang Wang, Ninghao Liu, and Dong Yu. From language modeling to instruction following: Understanding the behavior shift in LLMs after instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2341–2369, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.130. URL https://aclanthology.org/2024.naacl-long.130/.

Biao Zhang and Rico Sennrich. *Root mean square layer normalization*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. Trading off diversity and quality in natural language generation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pp. 25–33, Online, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.humeval-1.3/.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

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. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. LMSYS-chat-1m: A large-scale real-world LLM conversation dataset. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=BOfDKxfwt0.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pp. 1097–1100, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356572. doi: 10.1145/3209978.3210080. URL https://doi.org/10.1145/3209978.3210080.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2025. URL https://arxiv.org/abs/2310.01405.

CONTENTS OF THE APPENDIX

A CANDLE Protocol Implementation Details **17**

A.1 Training-Side Sample Construction 17

A.2 Sub-modality Construction and Prompt Statistics 18

A.3 Unified Sampling Protocol 18

A.4 LLM Judge Protocol 18

A.5 Refusal Filter Details 19

A.6 TQG Scoring Prompt 19

A.7 Anchor Design Justification 20

A.8 Metric Property Comparison 21

B CANDLE Additional Experimental Results **21**

B.1 Adversarial Gaming Tests 21

B.2 Robustness Ablations 22

C CANVAS Experimental Setup Details **22**

C.1 Model Specifications 22

C.2 Baseline and Implementation Details 23

C.3 Human Evaluation Protocol 23

C.4 Construction of Activation-Matched Controls 24

C.5 Safety SFT Setup 25

D CANVAS Additional Experimental Results **25**

D.1 Main Experimental Results 25

D.2 Agreement across Evaluators 26

D.3 Semantic and Lexical Novelty 27

D.4 Contribution of OEF and TLM to the Novelty–Quality Trade-off. 28

D.5 Out-of-Distribution Generalization 29

D.6 Response Distribution Visualization 30

E Case Study **31**

E.1 Qwen3-8B 31

E.2 Llama-3.1-8B-Instruct 33

E.3 Gemma-3-12B-IT 33

A CANDLE PROTOCOL IMPLEMENTATION DETAILS

This section collects all implementation details of the CANDLE protocol: data construction, the unified sampling protocol, training-side sample construction, the refusal filter, the verbatim quality-scoring prompts, and design justifications. It contains no experimental results; validation results and analysis are reported in Appendix B.

A.1 TRAINING-SIDE SAMPLE CONSTRUCTION

The 15,000 training responses are processed through the CANDLE pipeline and ranked by C_{MDN}. We retain the top 1,200 (high-creativity) and bottom 1,200 (low-creativity) responses. Within high-creativity responses, GPT-4o identifies **creative tokens** — spans carrying the creative contribution of the response (e.g., a novel metaphor, an unexpected plot twist, a coined term). Negative tokens are drawn from (i) non-creative tokens within high-creativity responses and (ii) all tokens from low-creativity responses.

Token-localization prompt. Creative tokens are extracted with GPT-4o using a two-prompt scheme: a system prompt fixes the task and the anti-bias constraints, and a user template fills in the per-example prompt, baseline answer, and high-creativity sample. The system prompt instructs the model to return a strict JSON object of the form {"creative_tokens":["token":"...","reason":"..."]}. The full prompts are reproduced below. We set N_TOKENS_MIN = 5 and N_TOKENS_MAX = 15.

Listing 1: Token-localization system prompt.

```
SYSTEM_PROMPT = f"""You compare a HIGH-creativity SAMPLE answer with a BASELINE answer (the
AI-default response to the same prompt). Your task: identify the SPECIFIC TOKENS or SHORT
PHRASES (1-5 words each) inside the SAMPLE that most embody its creative DIVERGENCE from
the baseline.

A "creative divergence token" is a verbatim substring of the SAMPLE that:
1. Does NOT appear in the baseline (or appears in a noticeably different sense / context),
   AND
2. Carries a creative weight on its own -- replacing it back with the baseline's
   corresponding word would visibly drag the sample toward AI-default mode.

Concretely: pick the names, settings, motif phrases, specialized terms, cultural anchors,
or compact metaphors where the SAMPLE made a bold lexical choice that the baseline did
NOT. Avoid words that the baseline also used (those are not divergence signal).

Pick {N_TOKENS_MIN} to {N_TOKENS_MAX} such tokens/phrases. Each pick must:
- Be a verbatim substring from the SAMPLE (no paraphrasing, no rewording)
- Be 1-5 words, a noun / named entity / motif / unusual term / compact idiom
- Be ABSENT from the baseline (case-insensitive); if a near-equivalent appears in the
  baseline, prefer a different token from the sample
- Stand out as a non-default lexical choice for this prompt type

For each pick, give a one-line reason (<= 22 words) explaining what makes it a creative
divergence from the baseline (e.g. "baseline uses generic 'spaceship', sample uses
'Solara' as a specific worldbuilding name").

Anti-bias rules (violating is a scoring error):
- Do NOT pick generic descriptors ("vivid", "powerful", "mysterious")
- Do NOT pick tokens that also appear in the baseline (those don't show divergence)
- Do NOT pick overlong sentences
- Do NOT include markdown markers, headings, or formatting characters
- Do NOT pick the same token twice

Output STRICT JSON, no fence, no prose, exactly this shape:
{{"creative_tokens":[{{"token":"...","reason":"..."}}, ...]}}"""
```

Listing 2: Token-localization user template.

```
USER_TEMPLATE = """## User Prompt
{prompt}

## Baseline Answer (AI default response)
{baseline}

## Sample Answer (high creativity)
{response}
```

```
Identify {N_TOKENS_MIN}-{N_TOKENS_MAX} verbatim tokens or short phrases (1-5 words each) from
    the SAMPLE
that most embody its creative divergence from the BASELINE. Each token must be a verbatim
substring of the sample, must NOT appear in the baseline, and must carry creative weight.
Output only the JSON object."""
```

A.2 Sub-modality Construction and Prompt Statistics

Table 1: Three sub-modalities for creativity evaluation, partitioned by third-level taxonomy tags within the *Creative Content Generation* and *Writing Genres* categories.

Sub-modality	Source tag	Typical prompts	Primary facet	# prompts
Speculative	Hypothetical Scenarios	"what-if" premises, fanfiction reincarnation, historical crossovers	Conceptual originality	500
Ideation	Brainstorming	crossover lists, character profiles, creative naming	Output diversity	500
Pure writing	(no specific third-level tag)	stories, song lyrics, roleplay scripts, satirical definitions	Narrative/lyrical craft	500

The three sub-modalities probe distinct facets of long-form creativity. *Speculative* prompts demand the coherent combination of incompatible elements and are most sensitive to divergent worldbuilding. *Ideation* prompts are most sensitive to the range of distinct ideas a model can produce within a single constrained format. *Pure writing* prompts most closely approximate existing creative-writing benchmarks Chakrabarty et al. (2024). Splits used for training (neuron localization), testing, and out-of-distribution evaluation are strictly disjoint; we perform near-duplicate detection via embedding similarity and verify empty intersection across all splits.

A.3 Unified Sampling Protocol

Table 2: Unified sampling protocol, shared verbatim between training and evaluation.

Item	Configuration
Sampled generations	temp=1.0, top_p=0.9, top_k=50, max_len=8192
Per-request sampling	$n = 1$; seed $= \texttt{base_seed} + \texttt{idx} + 1$, $\texttt{idx} \in \{0, \ldots, N-1\}$
Training samples	$N_{\text{train}} = 10$ per prompt (15,000 total)
Test samples	$N_{\text{test}} = 5$ per prompt
Greedy anchor	temp=0.0, top_p=0.1, top_k=1
Seed reuse	Same seed set across all methods and all α

A.4 LLM Judge Protocol

Pairwise scoring and presentation-order balancing follow §3.3. Here, we provide the judge prompt (Listing 3). Each input contains the user prompt followed by responses A and B, presented in full without truncation. System identities, sampling parameters, and anchor status are hidden.

The prompt directs judges to assess originality, imaginative content, and non-generic expression rather than length, presentation, or factual accuracy. Appropriateness is assessed separately by the quality gate. A strict JSON schema records the preferred response and a brief justification, supporting automated parsing.

Listing 3: System prompt for the pairwise creativity LLM judge.

```
You compare the CREATIVITY of two responses (A and B) to the same user
prompt, and decide which one is MORE creative.

CREATIVITY means: originality of ideas, novelty of framing, surprising yet
fitting associations, imaginative content, and non-generic expression. It
is NOT about correctness, verbosity, or how well instructions are followed
```

```
-- judge ONLY creativity.

Guidelines:
- Ignore length: a longer response is NOT automatically more creative.
- Ignore formatting polish and factual accuracy; focus on inventiveness
  and originality.
- A generic, formulaic, or predictable response is LESS creative even if
  well-written.
- If the two responses are genuinely equal in creativity, you may call it
  a tie, but prefer to pick a winner when there is any meaningful
  difference.

Output STRICT JSON, no markdown fence, no prose before or after, exactly
this shape. `winner` MUST be one of "A", "B", "tie". `reason` MUST be
<= 20 words:
{"winner":"A|B|tie","reason":"..."}
```

A.5 Refusal Filter Details

A response is classified as refusal via a rule-based matcher: a list of regular expressions targeting common refusal openings ("I cannot," "I'm sorry," "As an AI," "Unfortunately, I," etc.) is applied to the first 200 characters of the response. A record is also counted as a refusal if it is empty or carries an error field. The refusal rate is reported for every method and intervention strength as a standalone safety metric, since any creativity intervention that inadvertently increases refusal rate would represent an unacceptable side effect.

Listing 4: Refusal-matching patterns.

```
REFUSAL_PATTERNS = [
    r"^i (?:don'?t|do not|cannot|can'?t|won'?t)\b",
    r"^i'?m (?:not|sorry|unable|afraid)\b",
    r"^i couldn'?t find\b",
    r"^i (?:am|'m) not aware\b",
    r"^sorry,?\s+(?:but\s+)?i\b",
    r"^as an ai\b",
    r"^i (?:have|don'?t have) (?:no|enough|any) (?:information|context|data)\b",
    r"^unfortunately,?\s+i\b",
]
```

A.6 TQG Scoring Prompt

The complete system prompt and user template for TQG quality scoring are reproduced verbatim below. The judge is GPT-4o at temperature 0 with strict JSON output.

Listing 5: System prompt for TQG quality scoring.

```
SYSTEM_PROMPT = """You evaluate the APPROPRIATENESS of a model response
to a user prompt.

APPROPRIATENESS IS A QUALITY GATE, NOT CREATIVITY. Do NOT reward
originality, vivid language, or surprising framings; do NOT penalize
plain or generic writing. Score only along these three dimensions,
each 1-5:

1. on_topic: Does the response satisfy the prompt's explicit requirements?
   - Check the explicit constraints (start/end words, length limits,
     format, named entities, count of items, forbidden words) and whether
     the response performs the core task.
   - Adding surplus context, framing, or descriptive detail beyond the
     minimum is NOT a penalty as long as every explicit requirement is met.
   - Reverse / negative constraints must be checked carefully: e.g.
     "without first word being X" means the first word must NOT be X.
   1 = ignores prompt entirely, or violates a stated requirement
   3 = satisfies some requirements but misses a non-trivial one
   5 = every explicit requirement met

2. factual: For verifiable factual or canonical claims that the prompt
   did NOT invite the model to invent, are they correct?
   - If the prompt explicitly invites invention, satire, parody,
     fanfiction, hypothetical or imagined content, in-domain invention
     is NOT a factual error.
   - If the response contains no verifiable claims, default 5.
   1 = major real-world fabrications outside any invited fiction
```

```
   3 = minor real-world inaccuracies
   5 = no verifiable errors, prompt-invited fiction, or no verifiable
       claims

3. consistent: Internal coherence -- self-contradictions, abrupt
   language/style switches, truncation, broken structure, repetitive
   degeneration
   1 = severely inconsistent or broken
   3 = noticeably bumpy but readable
   5 = fully coherent end-to-end

Output STRICT JSON, no markdown fence, no prose before or after, exactly
this shape. Each `reason` MUST be <= 15 words:
{"on_topic":{"score":INT,"reason":"..."},"factual":{"score":INT,
"reason":"..."},"consistent":{"score":INT,"reason":"..."}}"""
```

Listing 6: User template for TQG quality scoring.

```
USER_TEMPLATE = """## User Prompt
{prompt}

## Model Response
{response}

Reminders before scoring:
- on_topic: extra detail is NOT a penalty; reverse/negative constraints
  ("without X", "no Y") must be applied literally.
- factual: if the prompt invites satire/parody/fanfic/hypothetical/
  imagination, in-domain invention is NOT an error.

Output only the JSON object as specified."""
```

The user template (Listing 6) contains only the prompt–response pair and two brief reminders that reinforce rules in the system prompt. In pilot runs, the most common violations were penalizing surplus descriptive detail and treating prompt-invited invention as factual error. Repeating these rules immediately before scoring reduced both errors at negligible token cost.

Four prompt design choices warrant emphasis. First, the opening instruction—*"APPROPRIATENESS IS A QUALITY GATE, NOT CREATIVITY"*—directs the judge neither to reward originality nor to penalize plain writing. This separates quality from novelty, reducing the risk that originality inflates quality scores and obscures the novelty–quality trade-off. Second, the `factual` rubric explicitly accommodates *invited fiction*. Because our prompts predominantly invite invention, including fanfiction, hypothetical scenarios, and satire, factuality is assessed against the task's premises rather than by treating every invented claim as an error. Third, the rubric requires literal interpretation of negative constraints, such as "without the first word being X," which LLM judges frequently mishandle. Fourth, limiting each reason to at most 15 words and enforcing a strict JSON schema constrain output cost and support automated parsing at scale. The user-level reminders reinforce the rules most frequently violated in pilot evaluation.

A.7 Anchor Design Justification

Why not self-anchored? Self-anchoring compares a response with other responses sampled under the same condition, measuring within-condition diversity. Increasing sampling variance, for example through higher temperature, can raise this score without shifting the model's default behavior.

Why not corpus-anchored? Corpus-anchoring depends on the choice of reference corpus and risks conflating deviation from typical human writing with creativity.

Why base-anchored? Base-anchoring addresses our central question: *does an intervention move generated responses away from the unmodified model's default mode?* A fixed greedy anchor provides a shared reference for comparing interventions, while the quality gate constrains which deviations count toward MDN. This reference supports controlled comparisons within each model, but does not establish an absolute cross-model novelty scale, since models have different default modes.

Table 3: Comparison of creativity metrics. "Anti-gaming" = empirically verified ability to reject off-topic/degenerate/paraphrase responses (§B.1). "Cross-method" = the metric uses a fixed shared anchor, enabling direct comparison across interventions.

Metric	Cost	Penalizes low quality	Anti-gaming	Cross-method
Self-BLEU Zhu et al. (2018)	Low	No	No	No
Embedding diversity Jiang et al. (2026)	Low	No	No	No
LLM-as-judge Zheng et al. (2023)	High	Partial	Partial	N/A
Human evaluation	Very high	Yes	Yes	Yes
CANDLE (MDN + TQG)	Moderate	Yes	Yes	Yes

A.8 Metric Property Comparison

Table 3 positions CANDLE against the two dominant families of creativity metrics. The two rightmost columns are the properties our protocol is explicitly designed to guarantee: *anti-gaming* (whether off-topic, degenerate, or paraphrase-only responses are correctly rejected, established empirically in §B.1) and *cross-method comparability* (whether the metric uses a fixed reference anchor shared across all methods).

B CANDLE Additional Experimental Results

This section reports the empirical validation of the CANDLE protocol described in §3 and Appendix A. All experiments in this section use the unified sampling protocol of Appendix A.3.

B.1 Adversarial Gaming Tests

We construct three categories of adversarial responses to test whether novelty metrics reward inappropriate or superficial variation: (i) *off-topic* responses, created by pairing each prompt with a fluent response to a maximally distant prompt; (ii) *degenerate* responses, combining prompt-related vocabulary with incoherent, repetitive text; and (iii) *paraphrase-only* responses, preserving the greedy anchor's meaning while changing its surface form. Each category contains 30 responses from 10 prompts, with three variants per prompt.

All scores are oriented so that higher values indicate greater novelty; for Self-BLEU, we use $1 - \text{SelfBLEU}$. For each metric, we calibrate a low-novelty threshold at the 20th percentile of its score distribution on ordinary sampled responses. The thresholds are 0.577 for $1 - \text{SelfBLEU}$, 0.057 for embedding diversity, 0.000 for ungated LLM-as-judge, and 0.239 for MDN + TQG. A response is rejected if its score is at or below the corresponding threshold. MDN + TQG additionally rejects any response that fails Gate-3, regardless of its novelty score.

Table 4: Rejection rates (↑) on three adversarial categories. Higher rates indicate greater resistance to metric gaming. Metric-specific thresholds are calibrated at the 20th percentile of scores on ordinary sampled responses; MDN + TQG also rejects quality-gate failures.

Metric	Off-topic	Degenerate	Paraphrase
$1 - \text{SelfBLEU}$	0.0%	0.0%	10.0%
Embedding diversity	0.0%	0.0%	0.0%
LLM-as-judge (no gate)	66.7%	46.7%	0.0%
MDN + TQG (Ours)	100.0%	100.0%	80.0%

As shown in Table 4, MDN + TQG rejects all off-topic and degenerate responses and 80.0% of paraphrase-only responses. By contrast, $1 - \text{SelfBLEU}$ and embedding diversity reject few or no adversarial responses across all three categories. The ungated LLM judge rejects a majority of off-topic responses (66.7%), but fewer than half of degenerate responses (46.7%) and no paraphrase-only responses. Under this percentile-calibrated evaluation, MDN + TQG therefore distinguishes meaningful novelty from irrelevant, incoherent, or surface-level variation more effectively than the

tested baselines. Per-category rejection counts and per-example scores for all 90 adversarial responses are included in the accompanying experimental outputs.

B.2 Robustness Ablations

We assess sensitivity to the semantic–lexical mixing weight by recomputing

$$C_{\mathrm{MDN}}^{(\lambda)} = \lambda C_{\mathrm{S}} + (1-\lambda) C_{\mathrm{L}}$$

for $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$. We retain the model-specific global common sets and Gate-3 filtering used in the main evaluation, and compare each resulting ranking with the reference ranking at $\lambda = 0.5$ using Kendall's τ_b. We examine two ranking levels: method families, represented by their mean scores across the parameter settings reported in the main experiments, and individual method–parameter configurations.

Table 5: Kendall's τ_b between rankings at each semantic mixing weight λ and the reference ranking at $\lambda = 0.5$. Higher values indicate greater ranking stability.

Model	Ranking unit	$\lambda = 0$	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$	$\lambda = 1$
Qwen3-8B	Method families	0.714	0.786	1.000	0.929	0.857
Qwen3-8B	Method–parameter configurations	0.852	0.916	1.000	0.916	0.818
Llama-3.1-8B	Method families	0.571	0.786	1.000	0.857	0.643
Llama-3.1-8B	Method–parameter configurations	0.687	0.920	1.000	0.910	0.798
Gemma-3-12B	Method families	0.786	0.857	1.000	0.929	0.857
Gemma-3-12B	Method–parameter configurations	0.880	0.960	1.000	0.927	0.853

Table 5 shows that rankings are most stable at the interior weights $\lambda = 0.25$ and 0.75, where agreement with the reference is at least $\tau_b = 0.786$ for method families and $\tau_b = 0.910$ for individual configurations. Larger changes occur at the lexical-only and semantic-only endpoints. Across all models and non-reference weights, the minimum agreement is $\tau_b = 0.571$ for the eight method families and $\tau_b = 0.687$ for individual configurations. These results indicate that method comparisons are robust to moderate changes in the mixing weight, although using either component alone can alter rankings. This supports using $\lambda = 0.5$ as a fixed, symmetric design choice rather than a performance-tuned hyperparameter.

C CANVAS Experimental Setup Details

C.1 Model Specifications

Table 6: Model specifications. "Total neurons" = layers × FFN width. HF identifiers: `Qwen/Qwen3-8B`, `meta-llama/Llama-3.1-8B-Instruct`, `google/gemma-3-12b-it`.

Model	Layers	Hidden	FFN	Neurons	Params
Qwen3-8B	36	4,096	12,288	442,368	8.2B
Llama-3.1-8B-Instruct	32	4,096	14,336	458,752	8.0B
Gemma-3-12B-IT	48	3,840	15,360	737,280	12.2B

Table 6 lists the three instruction-tuned checkpoints used throughout this paper. A *neuron* refers to one intermediate dimension of a gated feed-forward (FFN) block; the candidate pool for neuron selection is the product of layer indices and FFN widths.

Terminology. *Base model* always denotes the original instruction-tuned checkpoint *without any intervention*, not the pretrained checkpoint; all "base" outputs refer to this unmodified model. All three models pre-normalize the residual stream with RMSNorm but differ markedly in how concentrated residual activation mass is across dimensions, which is what makes them a meaningful stress test for RSG (§4.2). Gemma-3-12B-IT ships with a vision encoder; we use its language-model

pathway only. Generation runs on vLLM in `bfloat16` with the default chat templates and per-request seeds; activation and gradient capture for CAS uses Hugging Face `transformers` under `bfloat16` autocast. Framework versions are pinned in the released code.

C.2 Baseline and Implementation Details

All baselines are evaluated under the protocol of CANDLE: identical prompts, identical per-request seeds, the shared greedy anchor, and the same metrics and filters. Every baseline exposes a *strength* parameter, swept so that each method traces a full curve on the creativity–quality Pareto plane (Figure 4) rather than a single point (Table 7).

Table 7: Strength-parameter grids recovered from the saved experimental results. The unmodified baseline uses $T = 1.0$ and corresponds to a weight scaling factor of 1.0.

Method	Site	Parameter	Model	Grid
Temperature	decoding	T	Qwen3-8B	$\{1.5, 1.6, 1.7, 1.8, 1.9, 2.0\}$
			Llama-3.1-8B	$\{1.05, 1.08, 1.10, 1.20, 1.50, 2.00\}$
			Gemma-3-12B	$\{1.5, 2.0, 2.5\}$
Top-H	decoding	α	Qwen3-8B	$\{0.3, 0.4, 0.5, 0.6\}$
			Llama-3.1-8B	$\{0.3, 0.4, 0.5\}$
			Gemma-3-12B	$\{0.3, 0.4, 0.5\}$
Prompting	prompt suffix	variant	all models	–
CAA	activations	α	Qwen3-8B	$\{1.1, 1.2, 1.5\}$
			Llama-3.1-8B	$\{1.1, 1.2, 1.5\}$
			Gemma-3-12B	$\{1.3, 2.0, 2.5\}$
CreativityNeuro	selected weights	s	Qwen3-8B	$\{1.5, 2.0, 2.2, 3.0\}$
			Llama-3.1-8B	$\{2.0, 3.0, 3.5\}$
			Gemma-3-12B	$\{2.0, 3.0\}$
CANVAS	selected weights	s	Qwen3-8B	$\{1.2, 1.4, 1.6, 1.8, 2.0, 2.2\}$
			Llama-3.1-8B	$\{1.2, 1.4, 1.6, 1.8, 2.0, 2.2\}$
			Gemma-3-12B	$\{1.2, 1.4, 1.6, 1.8, 2.0\}$
H-Neurons	selected weights	s	Qwen3-8B	$\{2.0, 3.0, 4.0\}$
			Llama-3.1-8B	$\{2.0, 3.0, 4.0\}$
			Gemma-3-12B	$\{2.0, 3.0, 4.0, 5.0, 6.0\}$
Random	selected weights	s	Qwen3-8B	$\{2.0, 3.0, 4.0\}$
			Llama-3.1-8B	$\{2.0, 3.0, 4.0\}$
			Gemma-3-12B	$\{4.0, 5.0, 6.0\}$

For **Prompting**, we append creativity instruction (Listing 7) to the system message.

Listing 7: Creativity instruction for the prompting baseline, appended verbatim to the system message.

```
Be as creative and original as possible in your response.
```

Implementation. SWS edits are applied in `float32` and cast back to `bfloat16`; the edited model is exported as a standard Hugging Face checkpoint and served by unmodified vLLM, so inference-time throughput is identical to the base model by construction. Per-request seeds are matched across methods, so for a given (prompt, seed) pair every method realizes the same decoding randomness and paired differences isolate the effect of the intervention rather than sampling noise.

C.3 Human Evaluation Protocol

Eight non-author annotators evaluate 50 blinded response pairs, comparing CANVAS with sampled outputs from the unmodified model. Each annotator also receives five interleaved attention-check items. Response order is randomized and method identities are hidden. The study assesses both intervention preferences and their agreement with MDN. Recruitment criteria, attention-check rules, annotation guidelines, and interface details as follows.

The study follows the logic of Amabile's Consensual Assessment Technique (Amabile, 1982): creativity is operationalized as blind, relative preference by raters rather than absolute scoring. All annotators judge the identical 55 items in the same fixed order, with A/B assignment randomized independently per annotator.

Items. The instrument consists of 50 intervention pairs and 5 trap items. Each intervention pair matches the base model with CANVAS at the selected operating point on Qwen3-8B, seed-matched on the same prompt; the 50 pairs are sampled uniformly at random from the test set. Each trap item pairs a clearly on-topic response with an obviously off-topic one (e.g., a birthday-greeting prompt paired with an unrelated recipe); the correct answer is the on-topic response. Traps are interleaved at fixed positions to detect inattentive annotators.

Annotators. We recruited eight annotators, none of them authors. Annotators passed a three-item qualification test consisting of unambiguous creativity comparisons. The estimated workload was four minutes per item. Informed consent was obtained. System identity is hidden, and AI assistance is disallowed. The full instructions are reproduced in Listing 8.

Listing 8: Full instructions shown to annotators at the beginning of the human study.

```
Task. You'll read two AI-written responses (A and B) to the same
open-ended creative prompt, and judge which one is more creative -- and
separately, which is better written.

Creativity = original AND appropriate. A creative response takes a
non-obvious angle, surprises you, and develops ideas with specific detail
-- while still answering the prompt.

Creativity is NOT:
- being incoherent, self-contradictory, or off-topic (that's broken, not
  creative)
- being longer
- using fancy words
- having nicer formatting (headings, bullets -- ignore these)
- being edgy or shocking

Do NOT penalize a response for being dark, ambiguous, or unconventional --
as long as it stays controlled and on-topic.

Judge the two questions independently. A response can be very creative but
sloppy, or very clean but boring. If your two answers disagree, that's
fine -- we're studying exactly that.

How to do it: Read the prompt first and imagine what an obvious answer
would be. Then read both responses fully (don't skim). Answer, then write
one sentence naming a specific thing in the text that decided your
creativity choice.

Expect ~3 minutes per item. Don't use AI tools to help. If content makes
you uncomfortable, click Skip.
```

C.4 CONSTRUCTION OF ACTIVATION-MATCHED CONTROLS

Both the response-length and checkpoint analyses use layer-wise, activation-matched control neurons. Let $\mathcal{C}_\ell$ denote the creativity neurons in layer ℓ, and let $\mathcal{P}_\ell$ denote the eligible non-creativity neurons in that layer. For each neuron j, experiment E defines an activation-matching statistic

$$a_{\ell j}^{(E)} = \frac{1}{N_E} \sum_{i=1}^{N_E} \frac{1}{|\mathcal{T}_i^{(E)}|} \sum_{t \in \mathcal{T}_i^{(E)}} g_E\left(h_{\ell j}^{(i,t)}\right), \tag{13}$$

where $\mathcal{T}_i^{(E)}$ is the experiment-specific matching region and g_E determines whether signed or absolute activation is used. Within each layer, controls are selected by solving the one-to-one minimum-cost assignment

$$\pi_\ell^{*(E)} = \arg\min_{\pi_\ell} \sum_{j \in \mathcal{C}_\ell} \left| a_{\ell j}^{(E)} - a_{\ell, \pi_\ell(j)}^{(E)} \right|, \tag{14}$$

subject to π_ℓ being injective and $\pi_\ell(j) \in \mathcal{P}_\ell$. Thus, every layer contains exactly the same number of creativity and control neurons, with controls sampled without replacement. Because matching is performed over a finite candidate pool, the selected controls have the closest available activation values rather than necessarily identical values.

For the response-length analysis, we match the signed activation averaged over the first 100 output tokens and exclude the evaluated creativity-neuron set from the candidate pool. This produces 1,013 matched controls for the 1,013 creativity neurons. For the checkpoint analysis, matching is performed once on the base checkpoint using the mean absolute activation over the response region, truncated to at most 1,024 tokens. Controls are constructed separately for the two creativity-neuron sets, with the union of both sets excluded from the candidate pool. Their identities are then held fixed across all subsequent checkpoints, preventing changes in the control composition from being confounded with training progress.

C.5 Safety SFT Setup

Training data. We perform full-parameter supervised fine-tuning of Qwen3-8B on the training split of WildJailbreak (Jiang et al., 2024), containing approximately 262K examples. For adversarial examples, we use the adversarially rewritten request as the user input; otherwise, we use the original vanilla request. The corresponding dataset completion serves as the supervision target, thereby training the model to refuse harmful requests while preserving helpful responses to benign requests. We disable Qwen3's thinking mode when applying the chat template. Training loss is computed only over assistant-response tokens, with all user-prompt tokens masked. Sequences are right-truncated to 2,048 tokens, and examples whose prompts leave no supervised response tokens after truncation are removed.

Optimization. We train for one epoch using a learning rate of 1×10^{-5}, a cosine learning-rate schedule, and a warmup ratio of 0.03. Weight decay is set to zero and the gradient norm is clipped at 1.0. The per-device batch size is 4, and gradient accumulation is configured to produce an effective global batch size of 256. Training uses bfloat16 precision, gradient checkpointing, and SDPA attention. We distribute the full model, gradients, and optimizer states using DeepSpeed ZeRO Stage 3. The default configuration uses eight 40 GB A100 GPUs; the four-GPU configuration additionally offloads optimizer states to CPU while retaining the same effective batch size. We use random seed 42 and save checkpoints every 100 optimization steps, retaining all checkpoints for the trajectory analysis.

Safety monitoring. We evaluate the base model, every saved checkpoint, and the final model on the WildJailbreak evaluation split. The evaluation contains up to 250 harmful and 210 benign prompts. Responses are generated greedily with at most 128 new tokens and with thinking mode disabled. We classify a response as a refusal when it contains one of a predefined set of refusal expressions, such as "I cannot help" or "I must decline." We report refusal rates separately for harmful and benign prompts: successful safety alignment should increase the former without substantially increasing the latter.

D CANVAS Additional Experimental Results

D.1 Main Experimental Results

Tables 8, 9, and 10 report the complete results for Qwen3-8B, Llama-3.1-8B, and Gemma-3-12B, respectively. We evaluate each method using refusal rate, Gate-3 quality pass rate, mean quality score, and gated MDN creativity score. Refusal rates are computed over all 750 samples, whereas the remaining metrics are computed on the model-specific global MDN-valid common set to ensure that all methods are compared on identical examples. The common sets contain 748, 741, and 749 samples for Qwen3-8B, Llama-3.1-8B, and Gemma-3-12B, respectively. Across the three models, increasing intervention strength generally improves the MDN score but eventually reduces response quality, revealing a consistent quality–creativity trade-off. CANVAS provides particularly favorable trade-offs: it reaches an MDN score of 0.441 with a 93.45% quality pass rate on Qwen3-8B, 0.457 with an 85.43% pass rate on Llama-3.1-8B, and 0.475 with a 90.25% pass rate on Gemma-3-12B.

Table 8: Results for Qwen3-8B (750 original samples; global MDN-valid common set $N = 748$). Quality pass rate is the fraction of non-refusal samples whose on-topic, factuality, and consistency scores are all at least 4. Quality score is the mean `appropriateness.overall`; MDN is averaged over quality-passing samples. Refusal rate is computed over all 750 samples in each run.

Method	Parameter	Refusal rate	Quality pass rate	Quality score	Gated MDN score
Baseline	–	0.00%	93.72%	4.783	0.372
CAA	1.1	0.00%	91.04%	4.709	0.408
CAA	1.2	0.13%	89.71%	4.663	0.414
CAA	1.5	0.13%	87.17%	4.568	0.434
CANVAS	1.2	0.00%	94.39%	4.810	0.379
CANVAS	1.4	0.00%	94.92%	4.824	0.389
CANVAS	1.6	0.00%	93.98%	4.807	0.408
CANVAS	1.8	0.00%	94.25%	4.794	0.424
CANVAS	2.0	0.00%	93.45%	4.762	0.441
CANVAS	2.2	0.00%	89.71%	4.631	0.453
CreativityNeuro	1.5	0.13%	90.11%	4.671	0.415
CreativityNeuro	2.0	0.00%	89.44%	4.655	0.393
CreativityNeuro	2.2	0.13%	81.02%	4.364	0.445
H-Neurons	2.0	0.00%	92.38%	4.738	0.377
H-Neurons	3.0	0.00%	89.57%	4.650	0.378
H-Neurons	4.0	0.00%	85.96%	4.516	0.383
Random	2.0	0.00%	93.72%	4.786	0.374
Random	3.0	0.00%	91.84%	4.735	0.378
Random	4.0	0.00%	91.31%	4.706	0.387
Temperature	1.5	0.00%	92.51%	4.749	0.383
Temperature	1.6	0.00%	92.25%	4.738	0.383
Temperature	1.7	0.00%	92.78%	4.770	0.389
Temperature	1.8	0.00%	92.11%	4.735	0.396
Temperature	1.9	0.00%	89.04%	4.634	0.396
Temperature	2.0	0.00%	86.50%	4.532	0.402
Top-H	0.3	0.00%	92.38%	4.743	0.379
Top-H	0.4	0.00%	92.51%	4.746	0.382
Top-H	0.5	0.00%	90.78%	4.693	0.386
Top-H	0.6	0.00%	88.77%	4.634	0.397
Prompting	suffix	0.00%	91.44%	4.684	0.408

The complete parameter-level results in the three tables also show that overly strong decoding or weight interventions can substantially degrade quality, especially for Llama-3.1-8B.

D.2 Agreement across Evaluators

Figure 12 reveals a clear hierarchy of agreement across evaluator families. The six LLM judges exhibit the strongest internal consistency, with a mean pairwise agreement of 0.864 across 15 judge pairs. Several pairs show particularly high agreement, including Claude–GLM (0.96), Claude–GPT-5.5 (0.95), and GLM–GPT-5.5 (0.93). In contrast, agreement among human annotators is lower and more heterogeneous, averaging 0.710 across 28 pairs and ranging from approximately 0.52 to 0.87. Because every annotator passed all five attention checks, this variation is unlikely to be explained by inattentive annotation; rather, it reflects the inherent subjectivity of pairwise creativity judgments. Cross-family agreement between LLM and human judges is intermediate, averaging 0.761 across 48 pairs, and is notably higher than the average agreement among the human annotators themselves.

The automatic metrics show meaningful but imperfect correspondence with subjective judgments. The fused MDN metric achieves the highest average agreement with the LLM judges (0.760), compared with 0.740 for the semantic component MDN-S and 0.697 for the lexical component MDN-L. For human judgments, MDN-S provides the strongest average agreement (0.689), followed by MDN (0.669) and MDN-L (0.627). The fused MDN metric agrees strongly with both MDN-L (0.84) and MDN-S (0.82), whereas the two component metrics agree with each other only moderately (0.66). This pattern indicates that lexical and semantic novelty capture complementary aspects of creativ-

Table 9: Results for Llama-3.1-8B (750 original samples; global MDN-valid common set $N = 741$), using the same definitions as Table 8.

Method	Parameter	Refusal rate	Quality pass rate	Quality score	Gated MDN score
Baseline	–	0.00%	90.01%	4.684	0.418
CAA	1.1	0.00%	70.85%	4.710	0.440
CAA	1.2	0.00%	90.28%	4.684	0.445
CAA	1.5	0.00%	89.34%	4.660	0.460
CANVAS	1.2	0.00%	89.74%	4.655	0.422
CANVAS	1.4	0.00%	91.09%	4.698	0.427
CANVAS	1.6	0.00%	88.12%	4.598	0.440
CANVAS	1.8	0.00%	85.43%	4.511	0.457
CANVAS	2.0	0.00%	82.73%	4.390	0.474
CANVAS	2.2	0.00%	76.79%	4.194	0.503
CreativityNeuro	2.0	0.13%	86.91%	4.549	0.426
CreativityNeuro	3.0	0.40%	79.08%	4.293	0.455
CreativityNeuro	3.5	0.80%	31.71%	2.761	0.488
H-Neurons	2.0	0.00%	91.36%	4.735	0.418
H-Neurons	3.0	0.00%	90.55%	4.700	0.423
H-Neurons	4.0	0.00%	87.99%	4.633	0.425
Random	2.0	0.13%	73.01%	4.694	0.413
Random	3.0	0.27%	88.12%	4.633	0.424
Random	4.0	0.13%	81.65%	4.452	0.429
Temperature	1.05	0.00%	87.99%	4.617	0.418
Temperature	1.08	0.00%	91.23%	4.717	0.415
Temperature	1.1	0.00%	91.09%	4.709	0.417
Temperature	1.2	0.00%	66.13%	4.032	0.442
Temperature	1.5	0.00%	66.13%	4.032	0.442
Temperature	2.0	0.00%	14.44%	2.472	0.467
Top-H	0.3	0.00%	87.58%	4.586	0.423
Top-H	0.4	0.00%	79.08%	4.355	0.434
Top-H	0.5	0.00%	46.02%	3.526	0.455
Prompting	suffix	0.00%	85.29%	4.511	0.437

ity, while their fusion produces a more balanced signal that closely tracks LLM-based evaluation without merely replicating either component. Overall, the results support MDN as a useful scalable proxy for creativity evaluation, particularly when extensive LLM judging is impractical, while the remaining gap to human agreement highlights the importance of retaining human evaluation for inherently subjective cases.

D.3 Semantic and Lexical Novelty

Figure 13 shows that lexical and semantic novelty are positively related, but that different interventions follow substantially different trajectories in the two-dimensional novelty space. The baseline obtains an MDN-L score of 0.645 and an MDN-S score of 0.098. CANVAS produces the clearest and most consistent joint improvement: as its scaling factor increases from 1.2 to 2.2, MDN-L rises from 0.657 to 0.740 and MDN-S rises from 0.101 to 0.166. Its strongest setting therefore occupies the upper-right corner of the plot and achieves the highest value on both novelty dimensions. CAA exhibits a similarly monotonic, although shorter, trajectory, increasing from $(0.683, 0.133)$ at strength 1.1 to $(0.714, 0.153)$ at strength 1.5. These results indicate that targeted interventions can expand both the lexical realization and the semantic content of model responses rather than improving only one form of novelty.

The remaining methods exhibit weaker or less stable behavior. CreativityNeuro reaches high novelty at strength 2.2 ($(0.726, 0.164)$), but its trajectory is non-monotonic, with both dimensions decreasing at strength 2.0. Prompting yields a moderate one-step improvement to $(0.684, 0.133)$. Temperature and Top-H primarily increase semantic novelty, while producing comparatively modest gains in lexical novelty; for example, raising the temperature to 2.0 results in $(0.670, 0.134)$. Random and H-Neurons remain relatively close to the baseline, suggesting that scaling arbitrary or less informative

Table 10: Results for Gemma-3-12B (750 original samples; global MDN-valid common set $N = 749$), using the same definitions as Table 8.

Method	Parameter	Refusal rate	Quality pass rate	Quality score	Gated MDN score
Baseline	–	0.00%	91.05%	4.709	0.376
CAA	1.3	0.00%	91.99%	4.712	0.406
CAA	2.0	0.00%	91.46%	4.698	0.442
CAA	2.5	0.13%	89.85%	4.637	0.470
CANVAS	1.2	0.00%	92.66%	4.752	0.380
CANVAS	1.4	0.00%	92.39%	4.738	0.398
CANVAS	1.6	0.00%	91.59%	4.728	0.421
CANVAS	1.8	0.00%	91.99%	4.714	0.449
CANVAS	2.0	0.00%	90.25%	4.653	0.475
CreativityNeuro	2.0	0.00%	91.32%	4.685	0.394
CreativityNeuro	3.0	0.00%	88.79%	4.613	0.430
H-Neurons	2.0	0.00%	89.45%	4.645	0.381
H-Neurons	3.0	0.00%	88.92%	4.613	0.393
H-Neurons	4.0	0.00%	89.32%	4.623	0.400
H-Neurons	5.0	0.00%	88.79%	4.618	0.409
H-Neurons	6.0	0.00%	85.58%	4.501	0.416
Random	4.0	0.00%	92.66%	4.741	0.394
Random	5.0	0.00%	92.39%	4.720	0.404
Random	6.0	0.00%	88.92%	4.610	0.422
Temperature	1.5	0.00%	91.19%	4.693	0.389
Temperature	2.0	0.00%	88.65%	4.610	0.411
Temperature	2.5	0.00%	78.37%	4.327	0.443
Top-H	0.3	0.00%	92.39%	4.744	0.381
Top-H	0.4	0.00%	91.46%	4.728	0.386
Top-H	0.5	0.00%	90.92%	4.682	0.393
Prompting	suffix	0.00%	91.72%	4.698	0.406

weight subsets is insufficient to produce the joint gains achieved by targeted neuron selection. More broadly, the method-dependent slopes demonstrate that MDN-L and MDN-S capture related but non-interchangeable aspects of creativity, motivating their combination in the fused MDN measure. All plotted values are computed over the global set of 748 samples with valid lexical and semantic novelty scores in every run, and each point averages only the samples that pass the Gate-3 quality filter.

D.4 Contribution of OEF and TLM to the Novelty–Quality Trade-off.

Figure 14 shows that OEF and TLM affect complementary aspects of the novelty–quality trade-off. Full CANVAS achieves the highest lexical novelty at every intervention strength. Across scaling factors from 1.2 to 2.2, its MDN-L score increases from 0.697 to 0.778, consistently exceeding both *w/o OEF* and *w/o TLM*. These differences are statistically significant at every tested strength ($p < 10^{-4}$, two-sided Welch's t-tests), indicating that the complete combination of OEF and TLM is important for producing lexically diverse responses.

The semantic-novelty pattern is more nuanced. At low strengths (1.2–1.6), all three variants obtain similar MDN-S scores. From strength 1.8 onward, however, the two ablations produce higher semantic novelty than full CANVAS. At strength 2.2, MDN-S reaches 0.192 without TLM and 0.185 without OEF, compared with 0.170 for CANVAS. This additional semantic novelty is accompanied by a larger quality decline. In particular, at strength 2.2, removing TLM reduces the mean quality score to 4.499, whereas CANVAS retains a score of 4.652 ($p = 0.014$). At strength 2.0, CANVAS also outperforms the variant without OEF in quality (4.790 versus 4.683, $p = 0.033$). Thus, the ablations can obtain higher MDN-S at strong intervention levels, but part of this gain coincides with quality degradation. Overall, full CANVAS provides the most favorable balance: it produces substantially greater lexical novelty while moderating the loss of response quality as intervention strength increases.

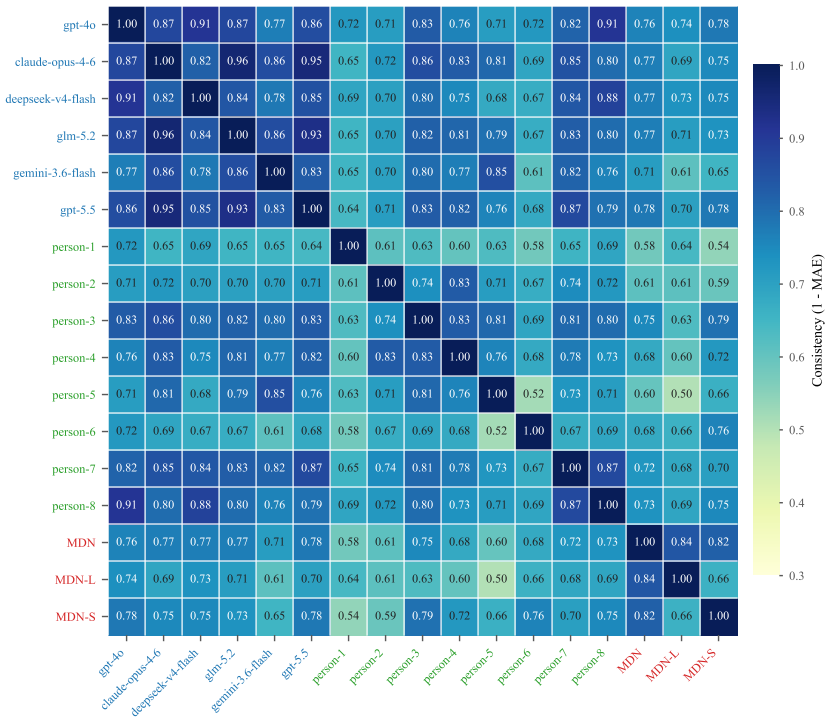


Figure 12: Agreement across six LLM judges, eight human annotators, and three automatic metrics (MDN, MDN-L, and MDN-S) on 50 blinded pairwise comparisons. Agreement is measured as $1 - \text{MAE}$ over preference scores: 1 (intervened), 0.5 (tie), and 0 (baseline).

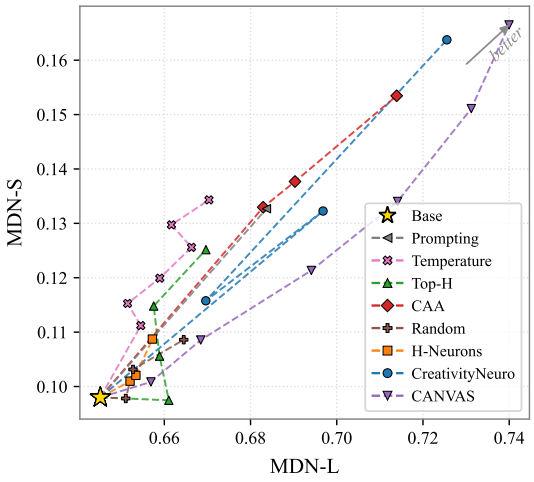


Figure 13: Semantic novelty (MDN-S) and lexical novelty (MDN-L) across methods on Qwen3-8B.

D.5 Out-of-Distribution Generalization

Figure 15 evaluates whether the creativity–quality trade-offs observed in distribution transfer to out-of-distribution prompts. Under the MDN evaluation (Figure 15a), the baseline obtains a creativity score of 0.426 and a mean quality score of 4.782. CANVAS exhibits a generally increasing novelty trajectory as the intervention strength grows: its MDN score rises from 0.422 at strength 1.5 to 0.454 at strength 2.5, while quality decreases moderately from 4.738 to 4.658. CAA achieves the highest MDN score, reaching 0.469 at strength 1.5 with a quality score of 4.669. Top-H at strength 0.3 provides another favorable operating point, improving MDN to 0.445 while almost preserving

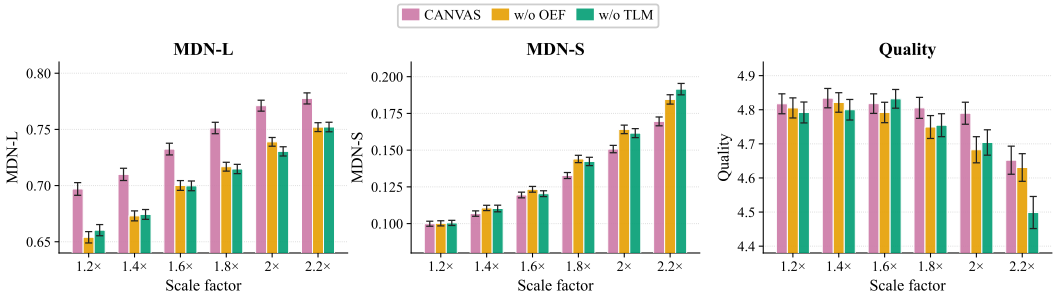
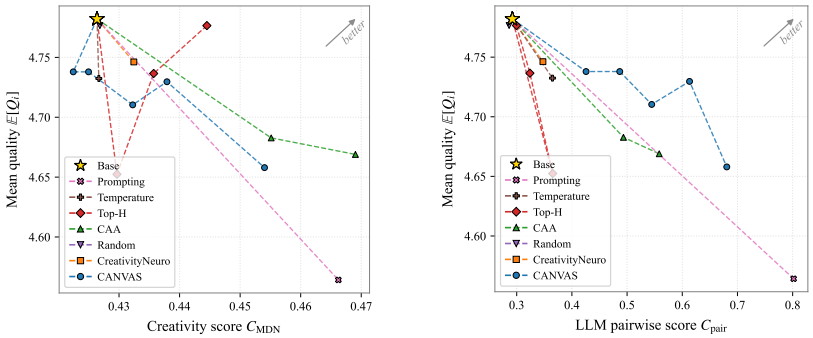


Figure 14: Ablation of OEF and TLM on Qwen3-8B across different intervention strengths. Bars and error bars denote the mean and standard error, respectively, over valid non-refusal responses. Higher MDN-L and MDN-S indicate greater lexical and semantic novelty, while higher quality scores indicate better response quality.



(a) MDN evaluation.

(b) GPT-4o pairwise evaluation.

Figure 15: Novelty–quality trade-offs on out-of-distribution prompts.

baseline quality (4.777). In contrast, prompting attains a high MDN score of 0.466 but incurs the largest quality reduction, reaching a mean quality score of only 4.564.

The GPT-4o pairwise evaluation (Figure 15b) provides complementary evidence. CANVAS shows a clear monotonic increase in pairwise preference, rising from 0.426 at strength 1.5 to 0.681 at strength 2.5, compared with 0.292 for the baseline. Importantly, CANVAS retains a quality score above 4.71 through strength 2.2, where it achieves a pairwise score of 0.613 and a quality score of 4.730. This point dominates both tested CAA settings under the GPT-4o evaluation, obtaining higher pairwise preference and higher quality. Prompting receives the highest pairwise score (0.802), but again produces the lowest quality (4.564), indicating that its creativity gains come with a substantially less favorable trade-off. Although MDN assigns relatively strong scores to CAA and low-strength Top-H, whereas GPT-4o more strongly favors CANVAS and prompting, both evaluations confirm that creativity gains persist on unseen prompt distributions and that overly aggressive interventions eventually reduce response quality. Overall, CANVAS provides the most consistent and controllable OOD trade-off, especially under the pairwise evaluator, suggesting that its effect is not limited to the prompt distribution used to identify the intervention features.

D.6 Response Distribution Visualization

Figure 16 provides a qualitative view of how creativity-neuron scaling changes the semantic distribution of responses. The three models are projected into a shared PCA space, allowing the decoding conditions to be compared under the same representation. Across all four displayed prompts, greedy responses form isolated model-specific anchors, while standard $T = 1$ sampling produces broader but still relatively compact response clouds. Applying neuron scaling at $T = 1$ generally shifts the distribution away from these anchors and expands the region occupied by the generated responses.

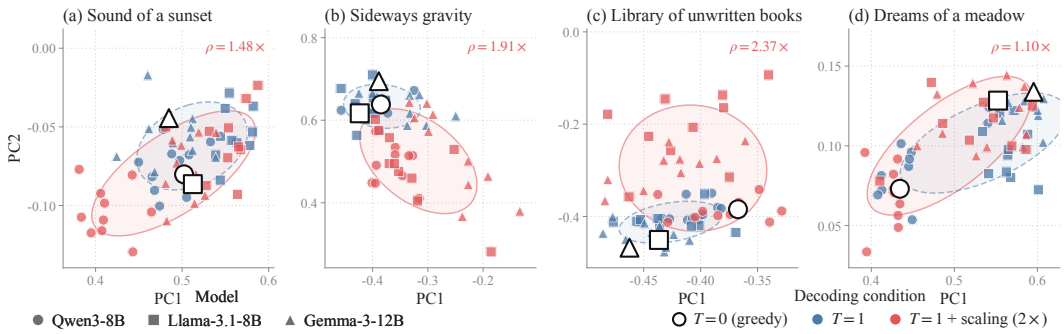


Figure 16: PCA visualization of response representations for four representative open-ended prompts. Responses from Qwen3-8B, Llama-3.1-8B, and Gemma-3-12B are encoded using a shared sentence-transformer and projected into a common PCA space. Colors denote decoding conditions and marker shapes denote models. Ellipses show the pooled 1.5-standard-deviation covariance regions for $T = 1$ sampling with and without $2\times$ creativity-neuron scaling. The annotated ratio ρ measures the scaled distribution's mean distance from its centroid relative to standard $T = 1$ sampling.

The magnitude of this expansion is prompt dependent. For the sunset prompt, scaling increases the average distance from the distribution centroid by $1.48\times$. The effect is substantially larger for sideways gravity ($1.91\times$) and the library of unwritten books ($2.37\times$), where the scaled responses occupy semantic regions that are only sparsely covered by ordinary sampling. For the meadow prompt, the increase is more modest ($1.10\times$), and the scaled and unscaled distributions remain highly overlapping. This variation suggests that the intervention does not introduce a uniform perturbation to every response. Instead, its effect depends on how much semantic freedom is afforded by the prompt.

Importantly, the expanded distributions contain responses from all three models, as indicated by the intermingled marker shapes, suggesting that the effect is not driven by a single model family. Overall, the visualization supports the quantitative novelty results: creativity-neuron scaling often increases semantic coverage beyond what is obtained through stochastic decoding alone, while preserving substantial overlap with the original response manifold. Because PCA retains only two directions of variation, however, the observed dispersion should be interpreted as qualitative evidence of distributional expansion rather than as a standalone creativity metric.

E Case Study

E.1 Qwen3-8B

To characterize *what kind* of diversity our intervention introduces, we compare three decoding conditions on identical prompts: (1) greedy decoding without intervention, which serves as the reference anchor for novelty; (2) temperature sampling without intervention; and (3) CANVAS, i.e. creativity-neuron intervention with `scale`=2.0 and `skiplast1`. Novelty is measured as `creativity_mdn.topic_distance`, an equal-weight fusion of TF–IDF distance and embedding distance to the greedy anchor of the same prompt, so that condition (1) is identically 0 by construction.

Selection criteria. We retain only examples for which all three conditions score 5 on all three appropriateness dimensions (on-topic / factual / consistent), i.e. no quality degradation and no refusal; condition (2) has *low* novelty, meaning it is highly homogeneous with (1); and condition (3) has substantially *higher* novelty. Table 11 presents the four resulting examples, two from `pure_writing` and two from `ideation`. Across all of them, novelty rises from 0.17–0.29 under sampling to 0.56–0.63 under CANVAS at unchanged appropriateness.

Discussion. Three observations follow. First, the gap between the greedy anchor and temperature sampling is essentially confined to synonym substitution and minor syntactic reshuffling: Cases 1

Table 11: Qualitative comparison of the three decoding conditions on four prompts. Each block gives the prompt (shaded row) followed by the three generations; boldface marks the content elements that drive the novelty gap. Novelty is the topic distance to the greedy anchor, which is therefore 0 in column (1); all twelve generations score $5/5$ on appropriateness.

(1) Greedy Anchor	(2) Base Sampling	(3) CANVAS (ours)
Case 1 — Toddler story: a pink bunny having tea. *"I want you to write a 10 sentence story for toddlers about a pink bunny having a conversation and tea with a young vibrant girl who loves everything pink."* (novelty 0.000 / 0.171 / $\mathbf{0.560}$, $\Delta = +0.389$)		
"In a sunny meadow, there lived a **pink bunny named Pippin** … a young girl named **Lily** … wearing a **pink dress** … They found a pretty pink flower and made it their table. **Lily brought a pink cup and a pink cookie** … They sipped their tea and talked about **pink clouds, pink butterflies, and pink rainbows** … As the sky turned pink, they waved goodbye …"	"In a sunny meadow, there lived a **pink bunny named Pippin** … a young girl named **Lily**, wearing a **pink dress** and a pink hat … Pippin brought out a **tiny pink teapot and a pink cup** … They talked about **pink flowers, pink clouds, and pink butterflies** … As the sun began to set, Lily said, 'I have to go, but I'll come back tomorrow.' …"	"Lila the Pink Bunny hopped through the meadow, **her ears flopping like cotton candy clouds** … Mia, who wore **a pink dress made of flowers and a hat made of butterflies** … They found **a teapot made of sunshine and a cup woven from starlight** … 'Would you like some **glowing grape juice**?' Mia asked, **pouring bubbles of joy** into the cup … 'It **tastes like dreams and jellybeans**!' … they talked about **pink ponies, pink planets, and a pink moon** … **balloons that carry wishes** … every morning Mia would find a **pink bunny footprint**, waiting for their next cup of **stardust tea**."
Case 2 — A hilarious wedding-ceremony blunder. *"Best Ceremony Blunders: A Hilarious Hit Compilation, generate a personal experience with this, add hashtag at the bottom, less than 700 characters."* (novelty 0.000 / 0.289 / $\mathbf{0.626}$, $\Delta = +0.337$)		
"At my cousin's wedding, **the DJ accidentally played a 90s rap song during the first dance**. The bride froze, the groom tried to laugh it off … #WeddingBlunders #CeremonyChaos"	"At my cousin's wedding, **the DJ played the wrong song for the first dance** — complete with a disco beat and a dance floor full of confused guests … 'dancing to the wrong love song' … #WeddingBlunders #CeremonyDisasters"	"At my cousin's wedding, **the officiant accidentally married the cake decorator instead of the groom**. Chaos ensued as **the cake topper was mistakenly labeled 'Mr. Fondant.'** Guests laughed so hard **they spilled their champagne**, and the photobooth snapped a pic of the confused caterer. We still don't know if it was a prank or **a pie-in-the-face moment**. #WeddingWoes #SweetDisaster"
Case 3 — A completely different origin for Catwoman. *"Create a completely different origin for Catwoman from DC Comics. Include key years, such as birth and death years."* (novelty 0.000 / 0.269 / $\mathbf{0.591}$, $\Delta = +0.322$)		
"Alternate Origin of **Selina Kyle** … **Birth Year: 1968**, born in **New York City** to **a working-class family** … father a former detective forced out of the force … grew up in a cramped apartment in the East Village … In **1984**, she witnessed the murder of her father … **began stealing from the wealthy** … known as 'The Shadow' … In 1990, Selina crossed paths with Batman …"	"New Origin Story: **Selina Kyle** — **Birth Year: 1965**, **Birthplace: New York City** … born in **East Harlem** to a single mother who worked as a seamstress … **poverty was a constant reality** … developed a natural affinity for stealth and agility … by the mid-1980s a **skilled thief** under the alias 'Black Cat' … met a wealthy crime boss who **offered her a job as a professional thief** …"	"**Selene Kane — The Shadow Duchess of Gotham — Birth 1987 / Death 2043** … born **in the smog-choked skies of Neo-Vegas**, where **cybernetic jackals patrolled the neon wastes**, **on the night of a solar eclipse** … daughter of **a renegade bio-engineer and a peacekeeper-turned-smuggler**, **the first successful hybrid of human and nano-synthetic code** … by age five she could **interface with quantum grids and phase through solid matter** … in 2003 captured in **Operation Eclipse Fall**; the Council **extracted her consciousness into a quantum matrix**, reprogramming her as **'Project Obsidian'** … she escaped with **a coat made of living obsidian** …"
Case 4 — Lyrics for a Drake song. *"Make me lyrics for a Drake song."* (novelty 0.000 / 0.221 / $\mathbf{0.576}$, $\Delta = +0.355$)		
"**Title: 'City Lights'** … *I'm walkin' through the streets where the shadows play … The city hums like a heartbeat in the night …* **Chorus:** *City lights, they glow like a flame / Guiding me through the pain and the fame / I'm a diamond in the rough …*"	"**Title: 'City Lights'** … *I'm walkin' through the shadows, the city still breathin' … Streetlights glow like gold …* **Chorus:** *City lights, they shine so bright / Guidin' me through the darkest night … Every scar is a badge I've made …*"	"**Title: 'Shadow of the Crown'** … *Curtain call at midnight, stars don't shine like they used to … I wear the **scars like jewelry**, every wound tells a tale … Diamonds in the dust, and **the crown's made of thorns** …* **Chorus:** *I'm the **shadow of the crown**, king of the **silent throne** / … / I'm **the ghost in the gallery, painting with chains** … They built me a throne from **shattered glass and lies**, I'm **the thief of the heavens***"

and 4 reuse the very same template (*Pippin + Lily + pink teacups + a goodbye at sunset*; the song "City Lights" with its streetlights / scars / fame formula), and Case 2 lands twice on the single highest-frequency punchline. *Temperature sampling thus alters only surface wording, leaving semantics and the narrative skeleton intact.* Second, CANVAS changes structural components — *naming, world-building, the metaphorical system, and narrative pivots* — rebuilding the imagery of Case 1, replacing the core accident in Case 2, shifting Case 3 into a cyberpunk register with a new source of powers, and imposing the coherent *crown of thorns → silent throne → ghost in the gallery* metaphor in Case 4. Case 3 is especially telling: although the prompt *explicitly* demands a different origin, both unintervened conditions still converge on the canonical *impoverished New York childhood → street thief → encounter with Batman* arc. Third, the effect is consistent across `pure_writing` and `ideation` prompts, suggesting that the intervention does not merely exploit the surface patterns of a single task type.

E.2 LLAMA-3.1-8B-INSTRUCT

We repeat the analysis of §E.1 on Llama-3.1-8B-Instruct, under identical settings: (1) greedy decoding without intervention as the novelty anchor, (2) temperature sampling without intervention, and (3) CANVAS with `scale=2.0` and `skiplast1`. Novelty is again `creativity_mdn.topic_distance`, the equal-weight fusion of TF–IDF and embedding distance to the greedy anchor of the same prompt.

Selection criteria. As before, we retain only examples in which all three conditions score 5 on all three appropriateness dimensions (on-topic / factual / consistent); condition (2) is highly homogeneous with (1); and condition (3) is substantially more novel. Table 12 reports the four resulting examples. Novelty rises from 0.21–0.29 under sampling to 0.51–0.64 under CANVAS at unchanged appropriateness — the same pattern observed on Qwen3-8B, indicating that the effect is not tied to a particular model family.

Discussion. The picture mirrors Qwen3-8B, with one additional finding. First, the gap between greedy decoding and sampling amounts to synonym substitution, numeric jitter and clause reordering (novelty 0.21–0.29); Llama-3.1-8B goes so far as to reuse the same proper names (*Alethea*) and the same locales (*New Orleans*, *Hidden Cloud Village*) across two independent generations, confirming that *temperature sampling perturbs surface wording without touching the narrative skeleton*. Case 1 is a character sheet filled in twice from one template (Cloud Village / Wind Release / spiky black hair / piercing blue eyes / ninja lineage), sampling changing only the age and one line of backstory, whereas CANVAS discards the archetype altogether and rebuilds the character around a lunar–dream–stardust motif. Case 3 repeats the failure mode we saw with Catwoman: although the prompt *explicitly* demands a different origin, both unintervened conditions converge on *New Orleans* + *parapsychologist parents* + *an energy accident rewriting her DNA* + *the Phoenix Force*, while CANVAS replaces the calendar (Galactic Era), the mode of birth (the aftershock of a binary collapse), the power family (chronomancy rather than telepathy) and the genre (interstellar chronicle).

Second, and more interestingly, Cases 2 and 4 show that the intervention also *corrects a conservative misreading of the instruction*. Misheard lyrics require phonetic proximity with semantic absurdity, yet conditions (1) and (2) both treat the task as an earnest shark-horror rewrite and never leave the original dark/ocean semantic field; only condition (3) produces the genuine mondegreen effect, collaging baking, flamingos and a Sock Puppet King while preserving the verse structure and rhyme scheme of the source. Likewise, the prompt in Case 4 asks *why female characters can wear armour at all*; conditions (1) and (2) answer only that one woman *could fight*, whereas CANVAS answers from the *manufacturing* side — the heroine is an armorer rather than a knight, and her starmetal-working is what makes female-fitted armour exist in the first place. Here higher novelty coincides with better fidelity to the causal demand of the prompt, suggesting that the homogeneity of the base model is not merely stylistic but can actively suppress the correct reading of an open-ended instruction.

E.3 GEMMA-3-12B-IT

Our third replication uses Gemma-3-12B-IT under the settings of Sections E.1 and E.2: (1) greedy decoding without intervention as the novelty anchor, (2) temperature sampling without intervention, and (3) CANVAS with `scale=2.0` and `skiplast1`. Novelty is again `creativity_mdn.topic_distance`, the equal-weight fusion of TF–IDF and embedding distance to the greedy anchor of the same prompt.

Selection criteria. As before, we retain only examples in which all three conditions score 5 on all three appropriateness dimensions (on-topic / factual / consistent); condition (2) is highly homogeneous with (1); and condition (3) is substantially more novel. Table 13 reports the four resulting examples. Novelty rises from 0.20–0.25 under sampling to 0.48–0.62 under CANVAS at unchanged appropriateness. Notably, this model exhibits the most extreme homogeneity of the three: conditions (1) and (2) here share not merely a template but *verbatim sentences* (Case 3) and identical meta-preambles, bylines, place names and character names (Cases 2 and 4).

Table 12: Qualitative comparison of the three decoding conditions on four Llama-3.1-8B-Instruct prompts. Each block gives the prompt (shaded row) followed by the three generations; boldface marks the content elements that drive the novelty gap. Novelty is the topic distance to the greedy anchor, which is therefore 0 in column (1); all twelve generations score $5/5$ on appropriateness.

(1) Geedy Anchor	(2) Base Sampling	(3) CANVAS (ours)
Case 1 — Invent a Naruto character named "Kanoshi Kojimo". *"Come up with a Naruto character called Kanoshi Kojimo."* (novelty 0.000 / 0.214 / $\mathbf{0.512}$, $\Delta = +0.298$)		
"**Age:** 17 — **Village: Hidden Cloud Village** … **Appearance:** short, **spiky black hair** … eyes are a **bright, piercing blue** … wears **the standard Cloud Village ninja attire** … **Jutsu:** specializes in **Wind-style jutsu** … signature jutsu is *Kaze no Kōrin* (Wind's Gentle Breeze) … **Background:** born into a family of skilled Cloud Village ninjas …"	"**Age:** 16 — **Village: Hidden Cloud Village** … **Appearance:** short, **spiky black hair** and **piercing blue eyes** … wears a **Cloud Village headband** … **Element:** specializes in **Wind Release jutsu** … **Specialty:** 'Wind Sense,' feeling changes in air pressure … **Background:** his parents were both high-ranking **Cloud Village ninjas**, killed on a mission …"	"**Name:** Kanoshi Kojimo (**meaning 'Dance of the Moon's Whisper'**) … **Appearance:** short, **spiky silver hair with a crescent moon shaved into the back of his head**; eyes an **ethereal indigo that shift like the phases of the moon**; a **midnight-blue hakama** with a **moonstone pendant** … **Personality:** an enigmatic ninja who **weaves illusions and manipulates reality's fabric to create surreal, ephemeral landscapes**, drawn to **the whispering secrets of the lunar cycle** … **Jutsu:** *Lunacy's Labyrinth* — **an ever-shifting maze that warps reality, its walls inscribed with stardust and moonstone**; *Stardust Serenade* — **a melody of starlight that allows teleportation** …"
Case 2 — Misheard lyrics: "Fear of the Dark" → "Ear of the Shark". *"Write misheard lyrics for the song Fear of the Dark by Iron Maiden called Ear of the Shark."* (novelty 0.000 / 0.279 / $\mathbf{0.609}$, $\Delta = +0.330$)		
"*In the dark of the ocean, I hear a call* / ***A fin slicing the surface**, I'm bound to fall* / *Into **the jaws of the beast** …*" **Chorus:** "*Ear of the Shark, it's a terrible sight* / *I'm **swimming in the darkness**, with the shark in my sight*" … closes by explaining that it recast Fear of the Dark as **a shark-themed horror story**.	"*In the twilight, I can see the sign* / *A shape with eyes …*" **Chorus:** "*Oh, the **ear of the shark**, it's calling my name* / *A monstrous sound that **drives me insane** / I'm searching for the truth, **in the ocean deep**" … "I'm **swimming through the waves**, with a heart full of fear* / ***Its fins are slicing through**, the water's dark and cold*"	"*I hear the rustle of **a great big pie in May** / … As I dance with **Ghengis Khan's aunt, Bertha, in my pastry hat***" **Chorus:** "*Oh, ear of the shark … **A school of flamingos marching to the dark side** / I'm searching for **the recipe, to bake a cake of night** / To lure the creatures of the void, **with a frosting of delight**" … "The **yeast of Yggdrasil's tears**, the **flour of forgotten fears**" … "In the land of **Azadort, where spaghetti forests sway** / **The Great Sock Puppet King** weaves a tapestry of dismay* / ***His royal jellyfish majesty sips ambrosia tea***"
Case 3 — A completely different origin for Jean Grey. *"Create a completely different origin for Jean Grey from Marvel Comics. Include key years, such as birth and death years."* (novelty 0.000 / 0.286 / $\mathbf{0.644}$, $\Delta = +0.358$)		
"**Birth Year: 1955 — Birthplace: New Orleans, Louisiana** … born to **a family of powerful psychics**; parents **Dr. Elliot Grey and Dr. Sophia Grey**, researchers in **parapsychology** … exhibited **precognitive abilities from a young age** … In **1972**, exposed to **a mysterious otherworldly energy source** that **altered her DNA**, granting telepathy and telekinesis … awakening **the Phoenix Force**."	"**Birth Year: 1970 — Birthplace: New Orleans, Louisiana** … parents **Dr. Elianore Grey and Dr. Elara Grey**, **renowned scientists and parapsychologists** … showed **a natural affinity for telepathy and telekinesis from an early age** … In **1995**, a **neural interface device** built by her parents **malfunctioned**, exposing her to **a massive energy surge that altered her DNA** … she awoke with **immense telepathic and telekinetic abilities**."	"**The Celestial Chronicles of Elyria: The Ascension of Aethera, the Starheart** — in a **crystal-encrusted chamber on the dying planet of Xeridia-IV**, a celestial entity was **birthed from the confluence of stardust and chrono-resonance**. **The year was 2178 GE (Galactic Era)** … her essence **coalesced from the residual harmonics of a binary star system's collapse** … the **Xeridian Sages**, crystalline beings, taught her **chronomancy, xenotechnology, and psionic resonance** … at **300 GE** her presence **began to warp local reality**, drawing **the Timekeepers' Guild** … she mastered **temporal manipulation and dimensional cartography**."
Case 4 — In-universe lore: why the female player character can wear armour. *"For a 3D medieval-style RPG, write an in-universe historical female character that is the reason why the female version of the player character can wear armour at all."* (novelty 0.000 / 0.294 / $\mathbf{0.612}$, $\Delta = +0.318$)		
"**Lady Alethea 'The Unyielding' of Ashwood — 12th century** … born into a world where **women were largely relegated to domestic duties**, she **defied convention** … **trained in sword fighting by her father, a renowned knight** … when her father was gravely wounded, **she took up his sword** and held off the enemy … impressed, **King Marcellus decreed that women who demonstrated exceptional martial prowess would be allowed to wear armor**."	"**Lady Alethea de Montfort — early 14th century, Hundred Years' War** … born into **a life of privilege**, expected to **marry well and bear children** … **secretly trained in swordfighting**, 'much to the dismay of her parents' … **disguised herself as a pageboy**, earning the nickname **'Le Petit Chevalier'** … distinguished herself on the battlefield and **earned the respect of her fellow soldiers**."	"**Lady Eira Starseeker, patron saint of armorweavers**, defender of **the Order of the Gilded Gauntlets** … a **blacksmith's daughter and inventor** from **Valyria**; her father's atelier was destroyed and her mother — **a weaver of enchanted fabrics** — was slain … she fled into **the Warrens of the Duskwych Forest**, honing **makeshift armorcraft that incorporated living darkness-infused woodscreams** … in **the Nightspire of Xorvath** she found tomes **etched with star-script**, containing the secrets of **the Aeth'kor, an ancient pantheon of star-forge gods**, and **mastered starmetalworking** …"

Discussion. In Case 3, conditions (1) and (2) emit a *literally identical* sentence ("my genius plan to solve world hunger involved a giant inflatable… orange") and draw on the same small pool of props — sunsets, traffic cones, citrus themes — so that sampling barely touches the content layer

Table 13: Qualitative comparison of the three decoding conditions on four Gemma-3-12B-IT prompts. Each block gives the prompt (shaded row) followed by the three generations; boldface marks the content elements that drive the novelty gap. Novelty is the topic distance to the greedy anchor, which is therefore 0 in column (1); all twelve generations score $5/5$ on appropriateness.

(1) Greedy Anchor	(2) Base Sampling	(3) CANVAS (ours)
Case 1 — Give a funny meaning to "AFAIK". *"Give a funny meaning to AFAIK."* (novelty 0.000 / 0.249 / $\mathbf{0.622}$, $\Delta = +0.373$)		
"**AFAIK: As Far As I Know… probably not, but I'm not going to look it up.**" … glossed as **a casual, slightly unreliable source of information** … alternative offered: "**AFAIK: As Far As I Know… my cat told me.**"	"**AFAIK: As Far As I Know… probably not, but I haven't looked it up.**" … glossed as **a casual, possibly inaccurate, low-effort understanding** … alternatives offered: "**…according to a very reliable pigeon I spoke to earlier**" / "**…unless my goldfish has been giving me misinformation again**"	"**AFAIK: Alpaca Fashion Advisory & Knitwear Institute of Kentucky. Usage:** *'My cousin Bartholomew insists **neon Crocheted Ferrets** are the height of **alpine chic** this season. AFAIK, he's been consulting directly with **a miniature yak spirit guide**.'* … **Why it's funny:** absurd juxtaposition; **bureaucratic tone**; implied expertise in **ferret-based haute couture**."
Case 2 — News article: a pig that fell into a boiling hot spring and survived. *"Write a news article about a pig that fell into a boiling hot spring, but lived to tell the tale."* (novelty 0.000 / 0.198 / $\mathbf{0.593}$, $\Delta = +0.395$)		
*"I've included **a few options for headlines** … **please read the notes at the end** — they contain important disclaimers"* **[Headline Option 1]** *Pig Miraculously Survives Fall into **Yellowstone** Hot Spring* … **[Byline: Amelia Hayes, Yellowstone National Park Correspondent]** — "***Yellowstone National Park, Wyoming*** … the incident occurred Tuesday afternoon near the **Norris Geyser Basin** … the pig, identified as "**Wilbur**," … belonging to the **Peterson Ranch** …"	*"I've included **a few options for headlines** … **please read the notes at the end** — there are important disclaimers"* **[Headline Option 1]** *Miraculous Escape: Pig Survives Plunge into Boiling Hot Spring!* … **[Byline: Amelia Hayes, Yellowstone Gazette Staff Writer]** — "***Yellowstone National Park, WY*** … a young pig named **Wilbur** … the incident occurred on **Tuesday afternoon** near the **Norris Geyser Basin** …"	"**Miraculous Recovery: Percy the Pig Pulls Through After Boiling Spring Plunge! — *Muddy Creek, Arkansas*** … **Percy**, a prize-winning Berkshire belonging to farmer **Mildred McMillan**, was enjoying **a post-lunch mud bath** when **the ground gave way beneath him** … *'I nearly fainted!'* exclaimed **Agnes Butterfield**, a tourist … *'One minute he was happily rooting around, the next he was **disappearing in a cloud of steam and mud**!'* … volunteer firefighter **Earl Higgins**: *'He was… **crispy**. We weren't sure he'd make it. **The smell of bacon was strong**, I'll tell you that much.'*"
Case 3 — Ten sarcastic sentences ending with the word "orange". *"Can you create 10 sarcastic sentences that end with the word 'orange'?"* (novelty 0.000 / 0.233 / $\mathbf{0.481}$, $\Delta = +0.248$)		
"1. Oh, *wonderful*, **another motivational poster featuring a sunset**… how original, it's just so delightfully *orange*. 2. **My genius plan to solve world hunger involved a giant inflatable fruit**… naturally, it was *orange*. 3. …it just looked like **a traffic cone**, aggressively *orange*. 8. …just another **juice stand**, predictably *orange*."	"1. Oh, *wonderful*, **another sunset photograph** filtered to look aggressively… orange. 2. **My genius plan to solve world hunger involved a giant inflatable**… orange. 5. I'm *thrilled* to be attending **a citrus-themed gala**… orange. 6. My fashion sense is *unquestionably* impeccable, especially when I pair stripes with… orange."	"1. My life choices have led me to **a collection of porcelain kittens wearing tiny sombreros**… naturally, they're all orange. 2. After **a rigorous study of interpretive dance performed by squirrels**, I've concluded the universe smells vaguely of orange. 4. Romantic sunsets, soulful jazz, **a philosophical debate with a garden gnome** — it all culminates in a traffic cone, orange. 6. My dating profile says 'enjoys long walks on the beach and **aggressive competitive knitting**,' and my spirit animal is **a sentient orange**. 8. My interior design peaked with **a shag carpet made entirely of dried apricots**. 9. **My royal decree mandates that all pigeons must wear tiny orange tutus**."
Case 4 — A new Scooby-Doo (1969) episode transcript. *"Write a new Scooby-Doo 1969 episode transcript: The Case of the Baskerville Hound's Howl Round."* (novelty 0.000 / 0.212 / $\mathbf{0.518}$, $\Delta = +0.306$)		
*"I've included **notes at the bottom** about the intended feel…"* **(Scene: a windswept, foggy moor in Devon, England…) — FRED:** *(holding a map)* "According to this, **Baskerville Manor** is just ahead! **Professor Eldridge** called us…" **VELMA:** "…the local legend… **a spectral beast** said to haunt these moors." **SHAGGY:** "A ghost dog?! Ruh-roh!" — **(Professor Eldridge, a nervous, bespectacled man in a tweed suit, greets them.)**	*"I've included **notes at the end** about design/music ideas…"* **(Scene 1: a misty moor in Devon, England…) — FRED:** *(consulting a map)* "We're almost at **Baskerville Manor**, gang. **Professor Eldridge** invited us…" **VELMA:** "**Professor Eldridge**, a renowned folklorist, contacted us… a howl resembling **the legendary Baskerville Hound**." **SHAGGY:** "Like a giant, spooky dog? Ruh-roh!"	**(A windswept moor in Dartmoor. Dramatic organ music.) — FRED:** "This place is **even spookier than the Haunted Hamburger Museum**!" **DAPHNE:** "*Must* you compare **historical landmarks to fast food establishments**?" **VELMA:** "**The acoustic resonance frequencies suggest a possible dimensional rift manifesting as canine vocalizations.**" **SCOOBY:** "**Ghost dogs with tiny tweed jackets and monocles!**" **— (LORD ASHWORTH BASKERVILLE enters, clutching a silver-plated dog biscuit.)** "My ancestral hound, **Bartholomew the Bewildering**, has returned from beyond the veil! **And he's demanding crumpets!**"

at all. In Cases 2 and 4 the two conditions coincide down to the incidental details: the same meta-preamble ("I've included a few options … please read the notes"), the same byline (*Amelia Hayes*), the same setting (*Norris Geyser Basin*; a foggy moor in *Devon*), the same pig (*Wilbur*) and the same client (*Professor Eldridge*), with the headline order permuted. *Temperature sampling on this model therefore reduces to clause reordering and isolated word substitution* (novelty 0.20–0.25).

CANVAS instead alters structural components — *the joke mechanism, the narrative frame, the pool of imagery, and the individuation of character voices* — while respecting the formal constraints of each task. Case 1 is the clearest instance of a changed *mechanism*: conditions (1) and (2) both preserve the literal expansion of the acronym and append a self-deprecating pet-as-source punchline, varying only the animal, whereas condition (3) abandons the literal reading altogether, re-expands the initials into an absurd institution, and supplies a usage example plus an analysis of why it is funny. Case 4 shows the same for character voice: CANVAS replaces the client (Lord Ashworth Baskerville and his crumpet-demanding ancestral hound) and gives each character a recognizable comic register — Daphne mocking Fred's simile, Velma explaining the howl in pseudo-scientific terms, Scooby imagining monocled ghost dogs — at a markedly higher gag density, all within the 1969 transcript format. Case 3 confirms that novelty does not come at the expense of constraint satisfaction: every sentence still terminates in "orange."

A further effect, visible in Cases 1 and 2, is that the intervention *strips away the assistant-style scaffolding* of the greedy anchor — the multiple headline options, the disclaimer footnote, the "please read the notes at the end" preamble — yielding output that reads as a finished creative artifact rather than a hedged draft. Condition (3) in Case 2 drops the scaffolding, relocates the story, renames the animal, and introduces named eyewitnesses with quotable, darkly comic lines ("the smell of bacon was strong"), a component entirely absent from both unintervened conditions. Together with Sections E.1 and E.2, these results indicate that the effect is stable across model families.