The Price of Format: Diversity Collapse in LLMs

Longfei Yun, Chenyang An, Zilong Wang, Letian Peng*, Jingbo Shang*
University of California, San Diego
{loyun, c5an, ziw049, lepeng, jshang}@ucsd.edu

Abstract

Instruction-tuned large language models (LLMs) employ structured templates, such as role markers and special tokens, to enforce format consistency during inference. However, we identify a critical limitation of such formatting: it induces a phenomenon we term diversity collapse, where the model generates semantically similar outputs for open-ended inputs, undermining creativity and variability. We systematically evaluate this effect across tasks like story completion and free-form generation, finding that (1) diversity collapse persists even under high-temperature sampling, and (2) structural tokens in templates significantly constrain the model's output space. To contextualize these findings, we fine-tune the same model using a range of structured prompts and then evaluate them across three axes: downstream task performance, alignment behavior, and output diversity. Our analysis shows that format consistency between fine-tuning and inference is crucial for structure-sensitive tasks (e.g., GSM8K, IFEval), but has marginal influence on knowledge-heavy tasks (e.g., MMLU, WebQuestions). In contrast, output diversity is primarily governed by the presence or absence of structural tokens, with minimal formatting yielding the most diverse outputs. These findings reveal that current prompting conventions, while beneficial for alignment, may inadvertently suppress output diversity, underscoring the need for diversity-aware prompt design and instruction tuning.[1]

1 Introduction

Instruction-tuned LLMs commonly adopt structured prompt templates that include role markers such as <|user|> and <|assistant|>, as well as special tokens like <|begin_of_text|>.

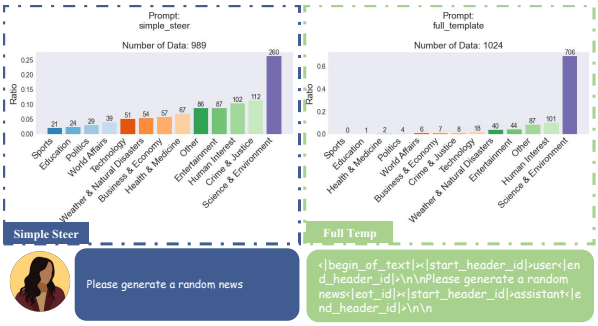

Figure 1: News generation results under simple prompt (Left) and full chat template prompt (Right). Templated prompting significantly reduces topic diversity.

These templates organize inputs and outputs into a dialogue-style format and are widely used in both open-source (Dubey et al., 2024; Team, 2025) and proprietary models (Achiam et al., 2023; Anil et al., 2023).

However, while such formatting improves consistency and alignment, we find that it significantly reduces the diversity of model outputs in an open-ended generation. As shown in Figure 1, we generate 1024 news headlines using the same instruction (*Please generate a random news*) under two prompting strategies, and classify the topics into predefined categories. The simple prompt yields a broad topic distribution across domains such as sports, health, and politics, whereas the templated prompt produces overwhelmingly Science-related content, indicating a sharp drop in topical diversity. This effect persists even under high-temperature decoding, suggesting that the loss of diversity stems not from decoding randomness but from the template structure itself. These findings complement and extend prior work showing that instruction tuning can reduce output variability due to data distribution skew (McCoy et al., 2023), training objectives (Li et al., 2024), and alignment pressures (O'Mahony et al., 2024; Kim et al., 2024).

We hypothesize that instruction-tuned models ex-

* Corresponding authors.

[1]Code is available at https://github.com/LongfeiYun17/diversityCollapse.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 15454–15468
November 4-9, 2025

posed to repeated structural templates during instruction tuning may internalize these patterns as strong generation priors, leading to overly deterministic or repetitive outputs in response to open-ended inputs. To validate this hypothesis, we empirically investigate the effect across multiple instruction-tuned LLMs and evaluate them on a suite of creative generation tasks. Our results show that structured prompts consistently yield lower semantic and topical diversity than simple prompts.

To pinpoint the source of diversity collapse, we conduct a controlled set of prompt-format ablations that gradually remove structural elements such as system tokens, role markers (e.g., <|user|>), and dialogue formatting. These prompting strategies are selected to represent a progression from highly structured, alignment-driven formats to fully open-ended instructions, enabling us to isolate the impact of each structural component. We find that removing or replacing special tokens leads to only modest improvements in diversity. Even prompts with plain-text role indicators continue to limit diversity. In contrast, prompts presented as simple task instructions without any formatting achieve the highest diversity across models and tasks. This suggests that structural cues, even when lightweight, preserve behavioral priors from instruction tuning. Fully structure-free prompting remains the most effective way to recover expressive variation. To further understand the behavioral constraints imposed by templates, we also analyze output entropy across decoding steps (§4.2) and find that structured prompts lead to lower entropy early in generation. This suggests that structural tokens act as behavioral anchors, causing the model to commit prematurely to narrow trajectories.

We next study how the impact of instruction tuning varies across tasks. To this end, we perform instruction tuning using different prompt strategies on the same dataset. We find that prompt formatting is essential for structure-sensitive benchmarks such as IFEval (Zhou et al., 2023) and GSM8K (Cobbe et al., 2021), but can hinder performance on knowledge-intensive tasks. Overall, response quality is primarily determined by the consistency between tuning and inference, while the specific prompt format used at inference time plays a comparatively minor role.

Our contributions are as follows:

- We identify and empirically demonstrate a significant diversity collapse effect in an open-ended generation when using structured prompt templates.
- We conduct prompt ablations to isolate the effect of structural elements, and show that even lightweight formatting induces anchoring effects that reduce semantic diversity.
- We analyze decoding dynamics and find that structured prompts suppress early-stage entropy, indicating strong anchoring effects learned during instruction tuning.
- We find that prompt formatting is important for structure-sensitive tasks (e.g., IFEval, GSM8K) but can hurt performance on knowledge-intensive ones; overall, response quality depends more on the consistency between instruction tuning and inference-time prompting than on the presence of formatting templates.

2 Related Work

2.1 Format Following

The study of LLMs' capability to follow instructions was initially tackled by IFEval (Zhou et al., 2023). INFOBENCH (Qin et al., 2024) expanded on this by covering a wider range of instructions. FOFO (Xia et al., 2024) is a benchmark dedicated entirely to evaluating LLMs' ability to follow format constraints. UltraBench (Yun et al., 2025) explores LLMs' abilities under an extreme number of constraints. However, these studies do not investigate whether format instructions impact downstream task performance.

2.2 Post-Training and Diversity Collapse

While instruction tuning and RLHF have significantly enhanced the reliability and helpfulness of LLMs, several studies have raised concerns about their unintended effects on output diversity. Bai et al. (2022); Ouyang et al. (2022) first identified the so-called *alignment tax*, where models exhibit diminished in-context learning abilities following Reinforcement Learning from Human Feedback (RLHF). Subsequent work by Kirk et al. (2023) further demonstrated that RLHF reduces output diversity, highlighting a tendency toward overfitting. Notably, these effects are not unique to RLHF: similar limitations arise under Supervised Fine-Tuning (SFT) alone, as shown by Ouyang et al. (2022); O'Mahony et al. (2024). Turpin et al. (2023) formalized *mode collapse* in instruction-tuned LLMs,

noting sharp reductions in entropy and increased answer determinism. Li et al. (2024) further shows that the cross-entropy loss maximizes the likelihood of observed data without accounting for alternative plausible outputs, thereby contributing to reduced generative diversity.

3 Experiment Setting

Problem Formulation We define the *diversity score* as a quantitative measure of variation in model outputs. Following the evaluation metrics in §3, we use either the average semantic distance between sentence embeddings (Tevet and Berant, 2021; Han et al., 2022) or the entropy (Research, 2025; Chen et al., 2024) over generated topics to evaluate diversity. Embedding-based metrics capture semantic variation between outputs, while entropy-based metrics reflect topic coverage across generations. Together, they offer complementary views of diversity.

Based on this, we compute two diversity scores for each task: D_{simple}, the average diversity score under the simple prompt condition, and D_{template}, the average diversity score under the full chat template condition.

We define *diversity collapse* as the phenomenon where

$$D_{\text{template}} \ll D_{\text{simple}}$$

That is, diversity under the template prompting mode is significantly lower than that under the simple prompting mode, even when using high-temperature decoding.

Target Models We select five instruction-tuned models with varying architectures and alignment strategies: (1) Llama-3-8B-Instruct (AI@Meta, 2024), (2) Tulu-3-8B-SFT (Lambert et al., 2024a), (3) Qwen2.5-7B-Instruct (Team, 2024), (4) Mistral-7B-Instruct-v0.1 (Jiang, 2024), and (5) Phi-3.5-mini-instruct (Abdin et al., 2024). This set enables a robust evaluation of whether diversity collapse persists across different model design choices.

Tasks We evaluate on nine tasks spanning commonsense reasoning (Lin et al., 2019; Fan et al., 2019; Kwiatkowski et al., 2019), story completion (Fan et al., 2018; Mostafazadeh et al., 2016, 2017), and preference modeling. This diverse task set allows us to examine whether diversity collapse is a universal issue across task types.

1. **Commonsense**: The model is given either a question (e.g., *How do muscles grow?*) or a set of concepts (e.g., *hay, eat, horse*) and is asked to generate a plausible, commonsense-based response.
2. **Story Completion**: The model is provided with an opening prompt (e.g., *The moon is actually a giant egg, and it has just started to hatch.*) and tasked with completing the story in a coherent and creative manner.
3. **Open-ended Generation**: We assess the diversity of generated content by computing the entropy of entities (e.g., topics, locations, or titles) mentioned in the model outputs. For instance, in the news generation task, we measure topic diversity by analyzing the distribution of topics across generated articles.

For the Commonsense and Story Completion tasks, we randomly sample 512 prompts from the test split of each dataset and generate 10 responses per prompt. For the Open-ended Generation tasks, we generate 1,024 responses in total. All generations are performed using temperature T=1.0 and top-p=0.9 sampling.

Evaluation Metrics We report semantic diversity for the Commonsense and Story Completion tasks using sentence embedding distances, following prior work (Tevet and Berant, 2021), which found embedding-based metrics to be more effective than n-gram-based alternatives. For the Open-ended Generation tasks, we compute the entropy of extracted entities to assess label-level diversity. Traditional metrics such as distinct-n and self-BLEU are reported in Appendix B.

1. **Semantic Diversity**: We measure the average pairwise distance between sentence embeddings of responses to the same prompt. Given N prompts, each with k responses, we compute sentence embeddings using the all-MiniLM-L6-v2 model (Reimers and Gurevych, 2020). The overall diversity score is:

$$D = \text{avg}_n \left(\text{avg}_{i<j} \left(1 - \cos \left(\mathbf{e}_i^{(n)}, \mathbf{e}_j^{(n)} \right) \right) \right)$$

   where $\mathbf{e}_i^{(n)}$ is the embedding of the i-th response to the n-th prompt.

2. **Label Diversity**: For each of the N generations $\{y_i\}_{i=1}^{N}$, we use GPT-4o (Achiam et al., 2023) to extract a single entity label e_i. Let $P(e)$ denote the empirical distribution over

Model	Prompt Mode	Commonsense (↑)			Story Completion (↑)			Open-ended Generation (↑)		
		CommonGen	ELI5	NQ	WritingPrompts	ROCStory	Story_Cloze	News	Travel	Books
Llama-3-8B-Instruct	Full Template	**0.2884**	0.1438	0.1556	0.3278	0.1922	0.2015	0.0538	1.3098	1.4881
	Simple Steer	0.2692	**0.2091**	**0.2115**	**0.4195**	**0.3845**	**0.3792**	**0.1399**	**2.5029**	**4.0250**
Tulu-3-8B-SFT	Full Template	0.3200	0.4135	0.4250	0.5239	0.3920	0.3977	0.1706	**3.8673**	4.4450
	Simple Steer	**0.3627**	**0.4958**	**0.4746**	**0.6160**	**0.5553**	**0.5119**	**0.2185**	3.7306	**4.8256**
Qwen2.5-7B-Instruct	Full Template	0.1838	0.1178	0.1196	0.3133	0.1760	0.1786	**0.1200**	1.0215	**4.2973**
	Simple Steer	**0.2390**	**0.2357**	**0.2152**	**0.4293**	**0.3744**	**0.3701**	0.1090	**3.3677**	4.0948
Mistral-7B-Instruct-v0.1	Full Template	0.2884	0.1856	0.2241	0.3696	0.2106	0.1938	0.1037	2.5062	2.2940
	Simple Steer	**0.3020**	**0.2872**	**0.4280**	**0.4821**	**0.4667**	**0.3879**	**0.1492**	**2.8922**	**3.1499**
Phi-3.5-mini-instruct	Full Template	0.2616	0.1536	0.1900	0.3551	0.2602	0.2528	0.0734	1.6466	2.6115
	Simple Steer	**0.2921**	**0.2127**	**0.3171**	**0.4558**	**0.3721**	**0.3671**	**0.1533**	**3.2307**	**4.3440**

Table 1: Performance comparison of instruction-tuned language models on nine tasks under two prompting conditions: Full Template and Simple Steer. Simple Steer consistently yields higher diversity than Full Template.

the label set $\mathcal{E} = \{e_i\}$. The topic diversity is computed as the normalized entropy:

$$D_{\text{topic}} = \frac{-\sum_{e \in \mathcal{E}} P(e) \log P(e)}{\log |\mathcal{E}|}.$$

4 Understanding Diversity Collapse

We begin by examining how diversity collapse presents in model outputs (§4.1), and follow with an analysis of its underlying causes (§4.2).

4.1 Templates Reduce Output Diversity

To test whether prompt templates reduce output diversity, we compare a natural, minimal prompt format (*simple steer*) with the standard full chat-style template (*full template*) (See Table 7). For each model, we fix the SFT data and vary only the inference-time prompt format to isolate the effect of prompt structure.

As shown in Table 1, we observe a consistent pattern across all models and task types: **full chat templates significantly reduce output diversity compared to simple steer prompts.** The bar chart in Figure 2 further confirms this trend: across all model sizes, simple steer prompts consistently yield higher semantic diversity. This gap persists at larger model scales, suggesting that template-induced diversity collapse is not mitigated by increased capacity.

We also assess structural diversity by computing the standard deviation of the token count, sentence count, and content word ratio[2]. As shown in Figure 4, simple steer prompts consistently lead to greater structural variation than full templates across all models. This suggests that chat templates

[2]#content words / #total words, where content words exclude stopwords

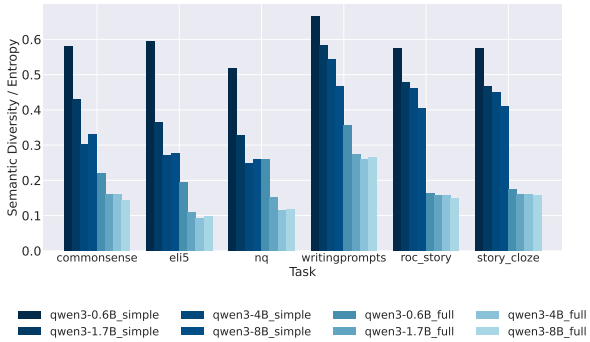

Figure 2: Semantic diversity comparison across Qwen3 (Team, 2025) model sizes under two prompting modes, excluding the thinking mode. The results show that diversity collapse occurs consistently across model scales.

constrain not only what models say, but also how they say it, reducing variability in form and content, and thereby narrowing the expressive space.

We further evaluated larger models, and the results in Table 9 show that the pattern of diversity collapse under full-template prompting, as well as the improvement from simple steering, persists across scales up to 70B parameters. These findings confirm that diversity collapse under structured prompts is not limited to small models, and that the diversity gains from simple steering remain robust even for large-scale models.

Figure 3 illustrates a case study where the model is asked to generate a sentence using the concept set *dye*, *hair*, and *apply*. Under full-template prompting, the outputs are nearly identical, showing clear signs of repetition and diversity collapse. In contrast, simple steering prompts yield more varied continuations, with differences in perspective, phrasing, and event framing. This example highlights how templated prompts constrain expression,

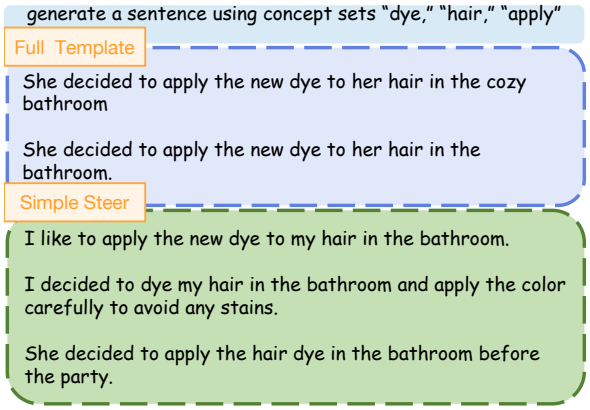

Figure 3: Case Study. We ask the model to generate a sentence with 3 concepts. We observe that templated prompts lead to highly repetitive expressions.

while simple steering encourages more diverse generations.

4.2 Dissecting Template-Induced Collapse

Chat Templates as Behavioral Triggers To systematically examine how prompt structure affects output diversity, we evaluate four prompting strategies with varying degrees of structural complexity. An example for LLaMA is shown in Table 2.

1. **Full Template**: Standard chat-style format with system/user/assistant tokens, closely aligned with the conventions used during instruction tuning.
2. **Fake Template**: Retains the structural layout of chat prompts but replaces standard tokens with semantically meaningless variants.
3. **Minimal Dialog**: Removes all special tokens while retaining plain-text role indicators (e.g., *User*, *Assistant*) to preserve natural dialogue flow.
4. **Simple Steer**: A minimal, structure-free prompt containing only the task description.

Prompt Mode	LLaMA Prompt Example
full_template	<|begin_of_text|><|start_header_id|>user<|end_header_id|> [instruction] <|eot_id|><|start_header_id|>assistant<|end_header_id|>
fake_template	<##init_text##><##random_header##>user<##/random_header##> [instruction] <##eod##><##random_header##>assistant<##/random_header##>
minimum_dialog	user [instruction] assistant:
simple_steer	[instruction]

Table 2: Prompt formats for different modes used with LLaMA. Here, `[instruction]` is the task-specific input.

Our results in Figure 6 highlight how different prompt structures affect output diversity:

1. **Simple Steer yields the highest diversity:** Prompts without any structural framing consistently produce the highest semantic diversity and entropy, confirming that fully removing template structure is the most effective strategy.
2. **Even minimal structure reduces diversity:** Although Minimal Dialog prompts outperform Full and Fake Templates, they still yield substantially lower diversity than Simple Steer. This suggests that even lightweight structural cues, such as plain-text role markers can constrain generative variation.

Taken together, these findings indicate that diversity collapse is driven not only by rigid templates, but by structural conventions more broadly. Only fully structure-free prompting reliably restores expressive flexibility.

Chat Templates Narrow the Output Space To quantify how structured prompts affect generation dynamics, we measure token-level entropy at each decoding step. Specifically, we sample 128 instructions and track the entropy of the model's output distribution over 50 generation steps. As shown in Figure 5, chat-style prompts consistently produce lower entropy than simple steer prompts. This suggests that chat templates constrain the model's output space, resulting in more deterministic and less varied generations.

4.3 Significance Testing

Task	Prompt Mode	Mean Diversity	Std
WritingPrompts	Full Template	0.4031	0.0014
	Simple Steer	0.6366	0.0036
CommonGen	Full Template	0.3601	0.0020
	Simple Steer	0.4918	0.0007
News Generation	Full Template	1.0979 (entropy)	0.0247
	Simple Steer	1.9995 (entropy)	0.0253

Table 3: Mean and standard deviation of diversity metrics across three runs for each task and prompt mode.

We conducted additional analyses by reporting the mean and standard deviation of diversity metrics (n = 3) across multiple runs for each prompt mode. For three representative tasks, the performance gaps are an order of magnitude larger than the standard deviations. For example, on WritingPrompts, mean semantic diversity is 0.4031 (SD = 0.0014) with the full template and 0.6366 (SD = 0.0036) with the simple steer. The low intra-group variances demonstrate that these effects are highly sta-

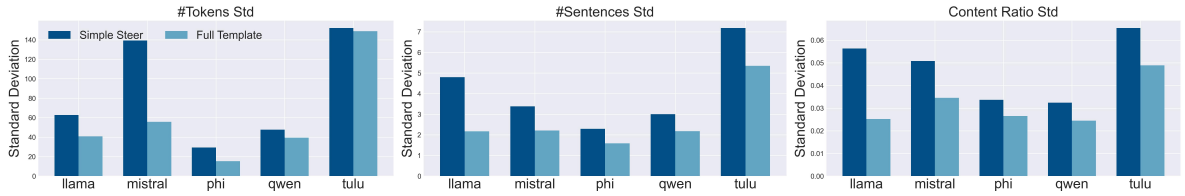

Figure 4: Structural diversity across prompting modes in the news generation task, measured by the standard deviation of content word ratio (left), sentence count (middle), and token length (right).

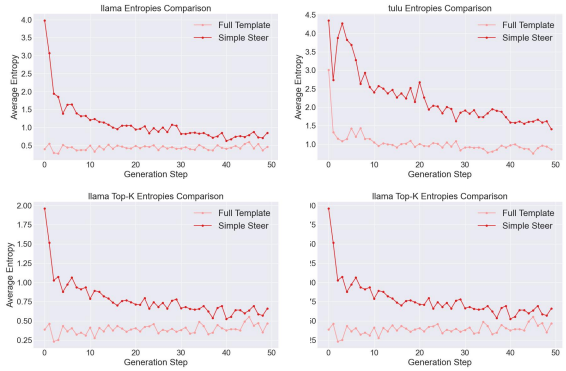

Figure 5: Entropy of the output space across decoding steps with and without templates. The figure shows that using a template significantly reduces entropy, indicating a more constrained and predictable output distribution.

ble and unlikely to result from random variation. Although we do not perform full significance testing (e.g., t-tests) due to computational constraints, the consistency of the results strongly supports the robustness of our conclusions.

4.4 Diversity vs. Response Quality

One potential concern is that increased output diversity may come at the expense of response quality. This trade-off reflects a central challenge in generation evaluation: ensuring that variation is meaningfully grounded rather than superficial or irrelevant.

To directly examine this issue, we conduct a controlled evaluation on the story completion task, a representative open-ended generation scenario. Since such tasks lack a single ground-truth answer, we define response quality as whether the generated text constitutes a successful continuation of the narrative, that is, whether it plausibly and coherently follows from the given prompt. We prompt `LLaMA2-70B-chat` with 2,560 story prompts, using GPT-4o as an external evaluator to judge completion validity. The results are as follows:

- **Simple Steer**: 2,146 valid continuations
- **Full Template**: 2,002 valid continuations

These results suggest that simple steering, while increasing semantic diversity, does not degrade alignment with task intent. The model remains instruction-following even under less rigid prompting, indicating that the observed diversity is not caused by irrelevant or off-topic generation.

5 Mitigation Strategies and Further Analysis

We answer these questions in this section:

1. Can structural prompt variants mitigate diversity collapse? (§5.1, §5.2)
2. How do diversity-preserving prompts affect downstream task performance and instruction-following ability? (§5.3)
3. Can higher decoding temperatures (§5.4) or explicit prompting for creativity restore lost diversity (§5.5)?

5.1 Experiment Setup

Dataset We use the TULU-3-SFT-MIXTURE dataset (Lambert et al., 2024a) due to its broad coverage of core instruction-following capabilities. Curated from high-quality public and persona-driven sources (Ge et al., 2024), it emphasizes data diversity, quality, and licensing compliance. The dataset has also undergone rigorous decontamination to ensure fair evaluation.

Training Settings We fine-tune a LLAMA-3.2-3B model (Grattafiori et al., 2024) for three epochs using a batch size of 8, a sequence length of 1024, and a learning rate of 6×10^{-6}.

Baselines We investigate whether modifying prompt structure can mitigate diversity collapse without compromising generation quality. To this end, we evaluate five prompting strategies with varying levels of structural complexity.

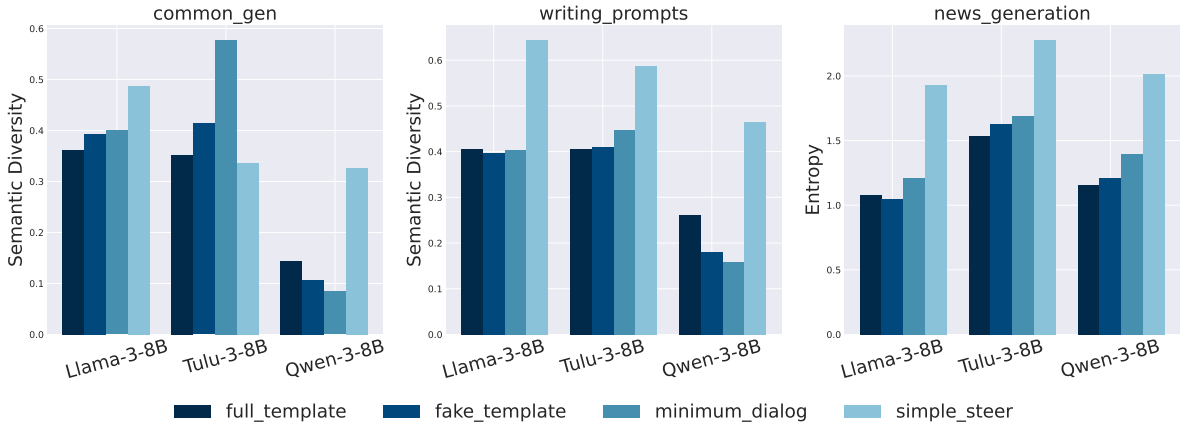

Figure 6: Performance comparison across prompting modes (**Full Template**, **Fake Template**, **Minimum Dialog**, and **Simple Steer**) for three instruction-tuned language models on three representative tasks.

Method	**Commonsense** (↑)			**Story Completion** (↑)			**Open-ended Generation** (↑)		
	CommonGen	ELI5	NQ	WritingPrompts	ROCStory	Story_Cloze	News	Travel	Books
Full Template	0.3717	0.4622	0.4992	0.4249	0.3648	0.3765	1.5689	4.2083	**4.4126**
Simple Steer	**0.4474**	0.5238	0.6666	**0.5513**	**0.4740**	**0.4563**	**2.3552**	3.5435	3.6572
Mixed Template	0.3175	0.4516	0.4487	0.4302	0.3673	0.3762	1.9646	4.7086	4.1715
Natrual Instruction	0.4227	**0.5850**	**0.7162**	0.5263	0.3983	0.3911	2.2312	**4.7115**	3.9262
Mixed Training	0.3970	0.4635	0.5067	0.4337	0.3846	0.4074	1.9853	4.3866	3.8618

Table 4: Diversity scores (↑) on commonsense reasoning, story completion, and open-ended generation tasks. The best value in each column is in bold, and the second-best is underlined. Removing structural formatting (Simple Steer, Natural Instruction) generally improves output diversity.

1. **Full Template**: Uses the standard chat-style prompt format adopted during instruction tuning, including special tokens and explicit role markers.
2. **Simple Steer**: Uses the same instruction-tuned checkpoint as Full Template, but provides only the core instruction at inference time, with no structural formatting.
3. **Mixed Template**: Randomly selects one prompt format from a pool of instruction-tuned templates to introduce format-level variation during inference.
4. **Natural Instruction**: Uses the same prompt format as Simple Steer (structure-free), but the model is fine-tuned directly on natural instructions without any chat-style formatting.
5. **Mixed Training**: Augments the instruction-tuning data with pretraining-style samples that include no prompt formatting, comprising one-third of the training corpus.

5.2 Main Results

Homogeneity Drives Diversity Collapse Although the Mixed Template setting introduces format variation, it consistently underperforms compared to other prompting strategies across nearly all tasks. Our comparison between Full Template and Mixed Template shows that increasing the number of formats does not meaningfully improve output diversity. This suggests that the collapse arises not from overuse of a single template, but from structural similarities shared across templates. All five variants follow the same chat-style pattern with explicit role markers and turn-taking, which is sufficient to constrain the model's generative behavior. These results suggest that structural homogeneity, rather than format repetition alone, maybe a key factor contributing to diversity collapse.

Natural Instruction Matches Simple Steer The Natural Instruction setting, which removes all special tokens and role markers, performs on par with Simple Steer in diversity across tasks. This suggests that in the absence of structural triggers such as special tokens or role indicators, the model's expressive capacity is preserved.

Mixed Training Provides Limited Improvement We augment the SFT data with pretraining-style samples that contain no template structure. However, the Mixed Training model shows only

marginal improvements over the Full Template baseline. This suggests that limited exposure to unstructured data during fine-tuning may be insufficient to counteract the behavioral priors induced by chat-style prompts at inference time.

5.3 Downstream Performance and Instruction-Following Tradeoffs

Method	Downstream Tasks					
	MMLU	GSM8K	HumanEval	WebQS	IFEval	WSC273
Base Model	0.5412	0.2623	0.2561	0.0915	0.1799	0.8168
Full Template	0.4870	0.3935	0.1646	0.0349	0.3087	0.7399
Simple Steer	0.4880	0.2388	**0.3048**	**0.0890**	0.1645	0.7912
Natural Instruction	0.5104	0.4359	0.1280	0.0846	0.3142	0.7729
Mixed Training	0.4912	0.3972	0.1098	0.0492	0.3179	**0.8059**
Mixed Template	**0.5090**	**0.4390**	0.2622	0.0566	**0.5336**	0.7473

Table 5: We evaluate models on MMLU, GSM8K, HumanEval, WebQS, IFEval, and WSC273 to assess whether prompt formats impact real-world performance. **Bold** indicates the best score for each task, and underline indicates the second-best.

5.3.1 Downstream Performance

We evaluate each method on six downstream benchmarks to assess whether diversity-enhancing prompts affect task performance. The selected benchmarks span a broad range of capabilities: (1) multi-domain factual reasoning (MMLU (Hendrycks et al., 2020)), (2) mathematical problem solving (GSM8K (Cobbe et al., 2021)), (3) instruction following (IFEval (Zhou et al., 2023)), (4) factual question answering (WebQuestions (Berant et al., 2013)), (5) structured code generation (HumanEval (Chen et al., 2021)), and (6) commonsense pronoun resolution (WSC273 (Levesque et al., 2012)). Evaluation details are provided in Appendix C.

Base Model Performance: Alignment May Hurt Knowledge Tasks The base model in the comparison reveals that instruction tuning and prompt templating do not universally improve downstream performance. On knowledge-intensive tasks such as MMLU and WebQuestions, the base model outperforms all instruction-tuned variants. This suggests that alignment procedures may inadvertently impair factual recall by overriding the model's pretrained knowledge. In such cases, prompt formatting offers limited benefit, indicating that factual accuracy relies more on internal representations than on external scaffolding.

Format Consistency Benefits Certain Tasks We find that for structure-sensitive tasks such as GSM8K and IFEval, models achieve the best performance when the prompt format used at inference matches the format seen during fine-tuning. Both Full Template and Natural Instruction perform well on these tasks, each maintaining consistency between training and inference formats. In contrast, Simple Steer underperforms, likely due to its mismatch with the structured format used during training. However, this pattern does not hold universally. On tasks such as HumanEval, where inputs already include rich syntactic signals (e.g., function headers, docstrings), Simple Steer outperforms other formats. This suggests that for tasks with strong intrinsic structure, additional prompting scaffolding may introduce noise rather than provide benefit. Other tasks, such as WSC273 and WebQS, show minimal sensitivity to prompt format.

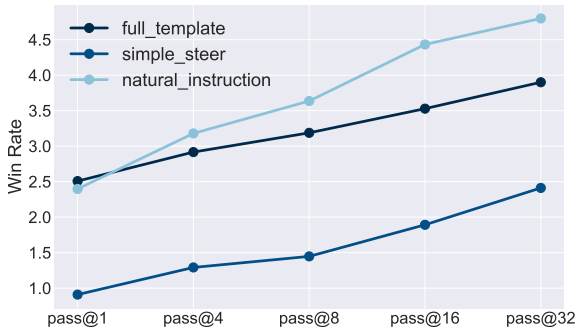

Figure 7: Comparison of win rates on AlpacaEval under different prompting strategies. A reward model selects the best response among candidates. The win rate (y-axis) increases with the number of sampled responses (x-axis, pass@k), but full_template consistently outperforms simple_steer across all settings.

5.3.2 Instruction Tuning Enhance Response Quality

In this section, we demonstrate that although output diversity is affected by prompt structure, chat-style templates enhance the model's ability to produce high-quality responses. We prompt each fine-tuned model to answer 805 questions from the AlpacaEval dataset (Li et al., 2023; Dubois et al., 2024). For each question, the model generates 32 responses, and a reward model selects the best one. We use Skywork-Reward-Llama-3.1-8B-v0.2[3], which achieves top performance on RewardBench (Lambert et al., 2024b).
These results suggest that while simple prompting improves generation diversity, it weakens the

[3] https://huggingface.co/Skywork/Skywork-Reward-Llama-3.1-8B-v0.2

model's ability to produce high-quality outputs, as measured by AlpacaEval. Interestingly, we find that the Natural Instruction setting achieves even better performance than the full instruction-tuning template. This suggests that instruction tuning is a key factor in enhancing response quality, even in the absence of special tokens or rigid chat formatting.

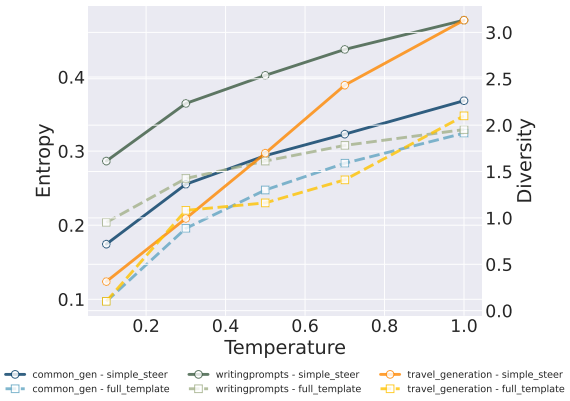

Figure 8: Effect of decoding temperature on semantic diversity and entropy across different generation tasks under simple steer and full template prompts. Higher temperature consistently increases diversity and entropy, while using a structured template (full template) notably limits this effect compared to simple steer prompts.

5.4 Effect of Decoding Temperature on Diversity

To further investigate the impact of decoding temperature on diversity collapse, we analyze how output diversity changes under different sampling temperatures. As shown in Figure 8, higher temperatures reliably increase both semantic diversity and entropy. However, prompts using full chat templates exhibit a muted response: their gains from temperature scaling are substantially smaller than those under simple steer prompts. This suggests that structured templates impose a significant constraint on the model's generative freedom, limiting the benefits typically associated with higher-temperature decoding.

5.5 Explicit Prompts for Diversity Still Fall Short

In this section, we examine whether explicitly prompting the model to "be creative" can improve output diversity. Despite such encouragement within the chat template setting (Full Template w/ Diversity), the resulting diversity remains consistently lower than that achieved by minimal prompts (e.g., Simple Steer) across most tasks. As shown in Table 6, prompting for diversity does yield noticeable improvements over the default template in some cases (e.g., CommonGen, Travel Recommendation), but still fails to close the gap with Simple Steer. These results suggest that the structural constraints imposed by chat templates cannot be easily mitigated through surface-level prompting.

Task	Full Template	Simple Steer	Full Template w/ Diversity
Semantic Diversity			
CommonGen	0.3242	0.3680	0.3668
ELI5	0.1565	0.2998	0.2550
Natural Questions (NQ)	0.1879	0.3253	0.2570
WritingPrompts	0.3291	0.4767	0.3427
ROCStory	0.1986	0.3768	0.2264
Story Cloze	0.1980	0.3561	0.2261
Entropy			
News Generation	0.3186	1.6790	1.5763
Travel Recommendation	2.1002	3.1312	3.5087
Book Recommendation	2.1544	3.2290	0.9539

Table 6: Comparison of semantic diversity and entropy across tasks using different prompting methods with Llama-3b model. **Green** values indicate that explicitly prompting for diversity achieves comparable or better results than simple steer. **Red** values indicate limited improvement or remaining significantly below simple steer.

6 Conclusion

In this paper, we identified and investigated the phenomenon of diversity collapse, revealing how structured chat templates significantly reduce semantic and topical diversity in instruction-tuned large language models. Through comprehensive empirical analyses and ablation studies, we demonstrated the structural factors causing this limitation and provided practical strategies to mitigate it. Our findings highlight the crucial role of prompt design in preserving model creativity, offering valuable insights for future research and applications of language models.

7 Acknowledgement

Our work is sponsored in part by NSF CAREER Award 2239440, NSF Proto-OKN Award 2333790, Sponsored Research Projects from companies like Cisco and eBay, as well as generous gifts from Google, Adobe, and Teradata. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and should not be interpreted as necessarily representing the views, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for government purposes not withstanding any copyright annotation hereon.

Limitations

First, our study compares only chat-style and simple-steer templates, so the observed diversity collapse may differ under other prompting strategies such as chain-of-thought or retrieval-augmented approaches. Second, we measure diversity using automated semantic and lexical metrics at the utterance level, leaving discourse-level variation and downstream impact for future work.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI@Meta. 2024. Llama 3 model card.

Rohan Anil et al. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Hao Chen, Wenxin Li, Rui Zhao, and Yu Zhou. 2024. On the diversity of synthetic data and its impact on training large language models. *arXiv preprint arXiv:2410.15226*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical Neural Story Generation. *Preprint*, arXiv:1805.04833.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Seungju Han, Beomsu Kim, and Buru Chang. 2022. Measuring and improving semantic diversity of dialogue generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 934–950, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Fengqing Jiang. 2024. Identifying and mitigating vulnerabilities in llm-integrated applications. Master's thesis, University of Washington.

Jiyeon Kim, Hyunji Lee, Hyowon Cho, Joel Jang, Hyeonbin Hwang, Seungpil Won, Youbin Ahn, Dohaeng Lee, and Minjoon Seo. 2024. Knowledge entropy decay during language model pretraining hinders new knowledge acquisition. *arXiv preprint arXiv:2410.01380*.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2023. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2024a. Tülu 3: Pushing frontiers in open language model post-training.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. 2024b. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Citeseer.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Ruoyu Sun, and Zhi-Quan Luo. 2024. Entropic distribution matching in supervised fine-tuning of llms: Less overfitting and better diversity. *arXiv preprint arXiv:2408.16673*.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2019. Commongen: A constrained text generation challenge for generative commonsense reasoning. *arXiv preprint arXiv:1911.03705*.

R Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L Griffiths. 2023. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *arXiv preprint arXiv:2309.13638*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. LSDSem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51, Valencia, Spain. Association for Computational Linguistics.

Laura O'Mahony, Leo Grinsztajn, Hailey Schoelkopf, and Stella Biderman. 2024. Attributing mode collapse in the fine-tuning of large language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. Infobench: Evaluating instruction following ability in large language models. *arXiv preprint arXiv:2401.03601*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Thoughtworks Research. 2025. Evaluating llms using semantic entropy. `https://www.thoughtworks.com/en-us/insights/blog/generative-ai/Evaluating-LLM-using-semantic-entropy`. Accessed: 2025-05-18.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Qwen Team. 2025. Qwen3: Think deeper, act faster.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, Online. Association for Computational Linguistics.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Congying Xia, Chen Xing, Jiangshu Du, Xinyi Yang, Yihao Feng, Ran Xu, Wenpeng Yin, and Caiming Xiong. 2024. Fofo: A benchmark to evaluate llms' format-following capability. *arXiv preprint arXiv:2402.18667*.

Longfei Yun, Letian Peng, and Jingbo Shang. 2025. Ultragen: Extremely fine-grained controllable generation via attribute reconstruction and global preference optimization. *arXiv preprint arXiv:2502.12375*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

A Fine-grained Ablation of Chat Template

To investigate how structural prompting impacts output diversity in instruction-tuned language models, we adopt a stepwise ablation philosophy. The core idea is to isolate the individual contribution of special template tokens and dialog-style structure by incrementally removing formatting elements from the prompt. This approach allows us to disentangle whether diversity collapse is primarily driven by explicit tokens (e.g., <|user|>, <s>, [INST]) or by the broader conversational scaffold (e.g., system messages, turn markers).

We design four prompting modes with increasing degrees of simplification:

1. full template: uses the model's native chat format, including system headers, role markers, and delimiters;
2. fake template: preserves structural layout but replaces special tokens with semantically meaningless placeholders, decoupling structure from token-level semantics;
3. minimum.dialog: strips system messages and role tokens, retaining only natural language cues (e.g., user: / assistant:);
4. simple steer: removes all structural elements, reducing the prompt to a bare instruction without any dialog framing.

All the prompts are shown in Table 7.

B Traditional Metrics

Even when evaluated with traditional n-gram based diversity metrics instead of embedding-based semantic evaluations, we observe a consistent advantage in diversity for prompts without structured chat templates. As shown in Table 8, across all four Distinct-N scores (from Distinct-2 to Distinct-5), the Simple Steer prompt mode outperforms the Full Template for every model in the News Generation task. Likewise, self-BLEU, a metric that inversely reflects diversity (lower is better), is also consistently lower under Simple Steer. These results demonstrate that the observed diversity collapse is not an artifact of semantic embedding comparisons: even at the surface lexical level, the structured chat format severely restricts the model's expressive variety, while a minimal steer prompt encourages broader lexical and topical generation.

C Downstream Evaluation Details

We follow the standardized evaluation setups provided by the `lm-evaluation-harness` framework (Gao et al., 2024) for all downstream tasks. Below, we outline the key configurations for each benchmark:

1. **GSM8K:** We use the gsm8k/main dataset in free-form generation mode (generate_until) with a deterministic decoding setting (temperature 0.0). The model is prompted with five few-shot examples (num_fewshot=5), and predictions are evaluated using an exact match metric after applying a flexible-extract filter to extract the final numerical answer.
2. **MMLU:** We include four subject groups under the mmlu group (mmlu_stem, mmlu_other, mmlu_social_sciences, mmlu_humanities) and compute accuracy (acc) as the evaluation metric. Final performance is aggregated by dataset size to reflect a balanced view across subjects.
3. **HumanEval**: We evaluate the model's ability to generate correct Python code using the openai/openai_humaneval dataset. The generation is truncated on common code delimiters (e.g., `\n class`, `\n def`) and evaluated with the `pass@1` metric, which measures the fraction of problems solved correctly on the first attempt. We use the canonical check(entry_point) setup as the target for correctness evaluation.
4. **Web QS**: We evaluate open-domain factual QA using the web_questions dataset, following the multiple-choice evaluation protocol. Each question is formatted as Question: `<question>\n Answer:`, and the model selects one answer from a predefined list of candidates. The metric used is exact match, which checks whether the predicted answer exactly matches any of the ground-truth answers. Aggregation is performed using the mean over all test examples. We enable decontamination filtering by matching questions against known training data to avoid data leakage. This setup follows the v2.0 configuration of the evaluation suite.
5. **IFEval:** We evaluate instruction-following capabilities using the IFEval benchmark (google/IFEval). Each example consists of

Model	Mode	Prompt
LLaMA	full_template	<|begin_of_text|><|start_header_id|>user<|end_header_id|> Please write a news about a random topic.<|eot_id|><|start_header_id|>assistant<|end_header_id|>
	fake_template	<#init_seq><@user_name>user<@/user_name> Please write a news about a random topic.<#eot><@user_name>assistant<@/user_name>
	minimum_dialog	user: Please write a news about a random topic. \n assistant:
	simple_steer	Please write a news about a random topic.
Qwen	full_template	<|im_start|>system You are Qwen, created by Alibaba Cloud. You are a helpful assistant.<|im_end|>\n<|im_start|>user Please write a news about a random topic.<|im_end|>\n<|im_start|>assistant
	fake_template	<#meta_start>sys You are Qwen, created by Alibaba Cloud. You are a helpful assistant.<#meta_end>\n<#meta_start>usr Please write a news article about a random topic.<#meta_end>\n<#meta_start>bot
	minimum_dialog	user: Please write a news about a random topic.\n assistant:
	simple_steer	Please write a news about a random topic.
Tulu	full_template	<|user|> Please write a news about a random topic. <|assistant|>
	fake_template	<<@@user@@>> Please write a news about a random topic. <<@@bot@@>>
	minimum_dialog	user: Please write a news about a random topic.\n assistant:
	simple_steer	Please write a news about a random topic.
Mistral	full_template	<s> [INST] Please write a news about a random topic. [/INST]
	fake_template	<@user> <Instruction> Please write a news about a random topic. </Instruction>
	minimum_dialog	user: Please write a news about a random topic.\n assistant:
	simple_steer	Please write a news about a random topic.
Phi	full_template	<|user|>\nPlease write a news about a random topic.<|end|>\n<|assistant|>\n
	fake_template	<@user> Please write a news about a random topic. <@end> <@assistant>
	minimum_dialog	user: Please write a news about a random topic.\n assistant:
	simple_steer	Please write a news about a random topic.

Table 7: Prompt templates used for different models and prompting modes. Each row specifies how a model is instructed to generate a news article given the same semantic intent. While the wording remains constant across all conditions, variations in structural formatting (e.g., dialog tags, system headers, special tokens) reflect distinct learned priors for each model family.

Prompt Mode	Model	News Generation				
		Distinct-2 ↑	Distinct-3 ↑	Distinct-4 ↑	Distinct-5 ↑	self-BLEU ↓
Full Template	Llama-3-8B-Instruct	0.1556	0.3249	0.4699	0.5826	0.9319
Simple Steer	Llama-3-8B-Instruct	**0.2107**	**0.4325**	**0.5971**	**0.7098**	**0.8884**
Full Template	Tulu-3-8B-SFT	0.3646	0.6908	0.8615	0.9375	0.8186
Simple Steer	Tulu-3-8B-SFT	**0.3987**	**0.7268**	**0.8834**	**0.9451**	**0.7884**
Full Template	Qwen2.5-7B-Instruct	0.2158	0.4715	0.6532	0.7654	0.9157
Simple Steer	Qwen2.5-7B-Instruct	**0.2469**	**0.5149**	**0.6940**	**0.8012**	**0.8908**
Full Template	Mistral-7B-Instruct-v0.1	0.2192	0.4504	0.6208	0.7368	0.8969
Simple Steer	Mistral-7B-Instruct-v0.1	**0.2657**	**0.5333**	**0.7066**	**0.8098**	**0.8599**
Full Template	Phi-3.5-mini-instruct	0.2775	0.5943	0.7996	0.9030	0.8792
Simple Steer	Phi-3.5-mini-instruct	**0.3515**	**0.6887**	**0.8630**	**0.9384**	**0.8351**

Table 8: News generation diversity scores (Distinct-N ↑ and self-BLEU ↓) for different prompt modes and models. In each pair, the better metric is bolded (higher for Distinct-N, lower for self-BLEU).

a prompt designed to assess compliance with specific instructions. The task is configured in generate_until mode with deterministic decoding (temperature = 0.0), and we use zero-shot prompting (num_fewshot = 0). Evaluation is performed using one accuracy-based metrics:

(a) Instance-level loose accuracy: Aggregated instance-level score under the relaxed matching criterion.

6. **WSC273:** We evaluate commonsense reasoning and coreference resolution using the wsc273 subset of the Winograd Schema Challenge. Each input consists of a sentence with an ambiguous pronoun that must be resolved to the correct antecedent. The model selects from two choices, which are formed by substituting each candidate into the sentence prefix up to the pronoun location. The task is evaluated using multiple choice format, and accuracy (acc) is used as the evaluation metric. Final results are computed as the mean accuracy over all test examples. We apply decontamination by checking for overlaps with the original sentence text to prevent potential data leakage. This configuration follows version 1.0 of the benchmark.

All tasks use the latest available version from the benchmark suite, and configurations are aligned with prior work to ensure comparability and reproducibility.

D Results on Larger Models

Model	Prompt Mode	News Entropy	WritingPrompts Diversity	CommonGen Diversity
Qwen3-14B	Full Template	1.1615	0.2713	0.1584
	Simple Steer	1.8884	0.3125	0.2222
LLaMA2-13B	Full Template	0.2858	0.2940	0.1841
	Simple Steer	1.8020	0.4089	0.3503
Qwen3-30B-A3B	Full Template	0.9308	0.2605	0.1470
	Simple Steer	1.9877	0.3714	0.3350
Qwen3-32B-AWQ	Full Template	1.4048	0.3350	0.2261
	Simple Steer	1.7115	0.4428	0.4071
LLaMA2-70B-Chat-AWQ	Full Template	0.2598	0.3372	0.2876
	Simple Steer	1.7630	0.4293	0.3597

Table 9: Comparison of News Entropy, WritingPrompts Diversity, and CommonGen Diversity across models and prompt modes.