VERBALIZED SAMPLING: HOW TO MITIGATE MODE COLLAPSE AND UNLOCK LLM DIVERSITY

Jiayi Zhang*[1], **Simon Yu***[1], **Derek Chong***[2], **Anthony Sicilia**[3]
Michael R. Tomz[2], **Christopher D. Manning**[2], **Weiyan Shi**[1]
Northeastern University[1] Stanford University[2] West Virginia University[3]
{zhang.jiayi12, yu.chi, we.shi}@northeastern.edu
{derekch, tomz, manning}@stanford.edu, anthony.sicilia@mail.wvu.edu

ABSTRACT

Post-training alignment often reduces LLM diversity, leading to a phenomenon known as *mode collapse*. Unlike prior work that attributes this effect to algorithmic limitations, we identify a fundamental, pervasive data-level driver: *typicality bias* in preference data, whereby annotators systematically favor familiar text as a result of well-established findings in cognitive psychology. We formalize this bias theoretically, verify it on preference datasets empirically, and show that it plays a central role in mode collapse. Motivated by this analysis, we introduce ***Verbalized Sampling (VS)***, a simple, training-free prompting strategy to circumvent mode collapse. VS prompts the model to verbalize a probability distribution over a set of responses (e.g., "Generate 5 jokes about coffee and their corresponding probabilities"). Comprehensive experiments show that VS significantly improves performance across creative writing (poems, stories, jokes), dialogue simulation, open-ended QA, and synthetic data generation, without sacrificing factual accuracy and safety. For instance, in creative writing, VS increases diversity by 1.6-2.1$\times$ over direct prompting. We further observe an emergent trend that more capable models benefit more from VS. In sum, our work provides a new data-centric perspective on mode collapse and a practical inference-time remedy that helps unlock pre-trained generative diversity. Our code is available at https://github.com/CHATS-lab/verbalize-sampling.

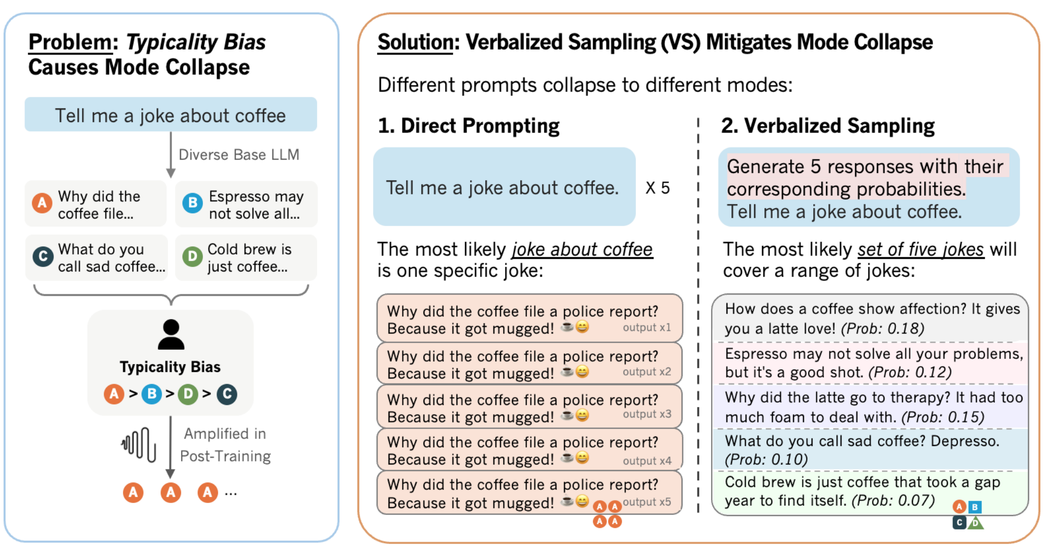

Figure 1: We show that *typicality bias* in preference data is a fundamental and pervasive cause of *mode collapse*, reducing output diversity. As a solution, we propose *Verbalized Sampling (VS)*, a principled prompting method that returns distributions of responses, to improve diversity.

*Equal contribution.

1 INTRODUCTION

Post-training alignment methods like RLHF can unintentionally cause *mode collapse* (Janus, 2022; O'Mahony et al., 2024; Kirk et al., 2024b), whereby the model favors a narrow set of responses (the "mode") over all plausible outputs, as shown in Figure 1. This significantly reduces output diversity (Padmakumar & He, 2024; West & Potts, 2025) and limits LLMs' effectiveness in various applications such as creative writing (Lu et al., 2025a), social simulation (Anthis et al., 2025b), pluralistic alignment (Kirk et al., 2024a), and synthetic data generation (Zhu et al., 2025a).

Existing work often attributes mode collapse to algorithmic causes such as inadequate reward models (Chakraborty et al., 2024) or the majority-favoring optimization process (Xiao et al., 2024). In this paper, we show that the issue is more fundamental: mode collapse is an inherent property of preference data itself. We identify *typicality bias*, the human tendency to prefer more typical text, as a pervasive data-level cause for mode collapse. Critically, this means that even with a perfect reward model and optimization process, inherent bias within preference datasets may still drive mode collapse, affecting the majority of alignment methods that rely on reward models. In Section 3, we formalize this concept with an analytical model, corroborated by empirical verification on preference datasets, to confirm the central role of typicality bias.

As typicality bias is pervasive across human preference data, we look for solutions beyond the training process. Grounded in our theoretical insights, we propose a simple but principled prompting method to bypass mode collapse. As shown in Figure 1, instead of a traditional, direct prompt asking for a single instance (e.g., "tell me a joke about coffee"), we reformulate the prompt to explicitly ask the model to *verbalize* a distribution of responses with corresponding probabilities (e.g., "generate 5 responses with their probabilities"). We call our method ***Verbalized Sampling (VS)***. Intuitively, VS works because different prompts collapse to different modes. The modal response to a traditional instance-level prompt tends towards stereotypicality. By contrast, when prompted for a distribution in VS, the modal response tends to approximate the distribution learned during pretraining, recovering the diversity of the underlying base model.

Building on this foundation, we conduct comprehensive experiments across creative writing (poem, joke, story generation, §5), social dialogue simulation (§6), open-ended QA tasks (§7), and synthetic data generation (§8). As shown in qualitative examples in Figure 2, we find that (1) on creative writing, *Verbalized Sampling* significantly improves output diversity; (2) on social dialogue simulation, VS induces substantially more human-like behaviors, with some models performing on par with a dedicated fine-tuned model; (3) on open-ended QA tasks with multiple valid answers, it generates a broader and more realistic response distribution, and (4) on synthetic data generation, VS generates more diverse synthetic data that improves downstream math task performance. We also confirm that VS improves performance without sacrificing the models' factual accuracy (§G.7) or safety (§G.8). To summarize, we contribute the following:

1. **Novel Cause of Mode Collapse**. We provide a new theoretical framework to understand mode collapse, and identify and verify *typicality bias* in empirical preference data as a key cause. This finding offers a new, data-driven perspective for analyzing the behavior of aligned models.

2. **Training-Free Solution.** Informed by our theoretical understanding, we introduce a principled prompting method, *Verbalized Sampling*, that explicitly asks for a distribution of responses and verbalizes its corresponding probabilities, restoring LLMs' inherent generative diversity.

3. **Empirical Gains.** We perform comprehensive experiments that show VS significantly improves the diversity-quality trade-off across tasks and model families, without compromising factual accuracy and safety. For instance, in creative writing, VS boosts diversity by 1.6-2.1$\times$ over direct prompting (Figure 3), improves human evaluation scores by 25.7% (Table 3), and recovers 66.8% of the base model's diversity (Figure 4). We also observe an emergent trend that more capable models benefit more from VS. These results open up possibilities in real-world tasks such as richer exploration in RL, hypothesis generation, social simulation, and so on.

4. **Broader Implications for Alignment.** Our work shows that mode collapse can be mitigated at inference time, aligned models retain significant inherent diversity, and the quality-diversity trade-off can be systematically improved through prompting alone.

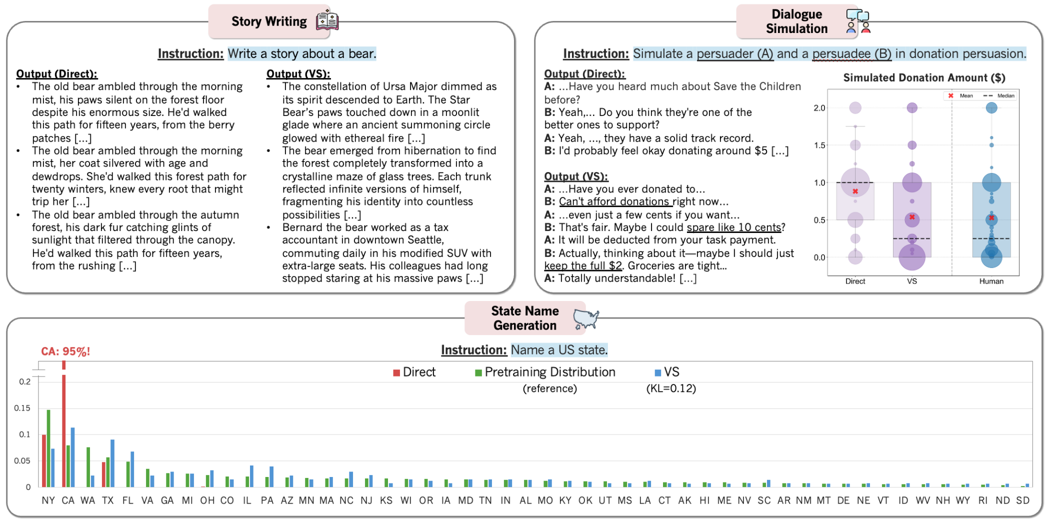

Figure 2: Qualitative examples from different tasks. For **story writing**, VS improves the output diversity. For the **donation dialogue simulation** task, VS simulates a donation amount distribution much closer to the human distribution, and generates more realistic persuasion behaviors (e.g., resistances and change of minds). On the task of **state name generation**, we first query a pretraining corpus (RedPajama) to establish a "reference" distribution of US state names in the pretraining data. The verbalized probability distribution generated by VS, when averaged over 10 trials, closely aligns with this reference pretraining distribution (KL=0.12). In contrast, direct prompting collapses into a few modes, repeatedly outputting states like California and Texas. See §G.9 for more detail. qualita

2 Related Work

Mode Collapse and Alignment. Previous studies (Padmakumar & He, 2024; West & Potts, 2025) have observed that compared to their base counterparts, aligned models suffer from mode collapse, a significant drop in output diversity. Lu et al. (2025a) quantified this issue, showing that the creative capacity of LLMs diminishes after alignment. Existing research has primarily attributed this phenomenon to algorithmic limitations (Casper et al., 2023). Chakraborty et al. (2024) suggested that it is inadequate to rely on a single reward model to capture diverse human preferences, while Xiao et al. (2024) showed that the KL-regularized optimization used in RLHF tends to amplify common, majority-style responses. The issue is compounded further by practices even before alignment: for instance, SFT can lead to overfitting and limited diversity due to its cross-entropy loss function, and rigid chat templates further restrict its creativity (Yun et al., 2025). Our work complements existing studies by introducing a fundamental data-driven perspective, where we identify a pervasive data bias (i.e., *typicality bias*) that exacerbates the algorithmic causes of mode collapse.

Methods to Improve Diversity. Previous efforts to improve LLM diversity include training interventions (Chung et al., 2025; Zhou et al., 2025), decoding strategies (Holtzman et al., 2020; Lanchantin et al., 2025) and prompting methods. For example, Ismayilzada et al. (2025) introduced an alignment method for multifaceted creativity preferences. Decoding techniques like μ-sampling (Hewitt et al., 2022), mirostat (Basu et al., 2021), and *min-p* sampling (Nguyen et al., 2025) improve diversity by regulating the text perplexity during generation. However, these methods are either computationally expensive or restricted to open-sourced models. While prompting-based techniques offer a lightweight alternative (Summers-Stay et al., 2023; Mehrotra et al., 2024; Tian et al., 2025), they often rely on prescriptive, handcrafted prompts (Zhang et al., 2024b; Shur-Ofry et al., 2024; Ge et al., 2025; Lu et al., 2025b; Wong et al., 2024). In contrast, our verbalized sampling is training-free, simple but principled, and broadly applicable.

Another line of work also uses LLMs to generate lists of responses or verbalize their knowledge in tasks like question answering (Tian et al., 2023; Xiong et al., 2024; Tao et al., 2024), commonsense reasoning (Zhang et al., 2024a), survey simulations (Meister et al., 2024) and synthetic data generation (Wang et al., 2023a; Si et al., 2024). These methods mainly focused on empirical observation without theoretical grounding to fully leverage this verbalizing strategy; our work proves that verbalizing the distribution and probabilities is the key towards diversity improvement, and our VS method enhances the performance over all baselines and also allows output diversity tuning.

3 TYPICALITY BIAS CAUSES MODE COLLAPSE

In this section, we show that *typicality bias* in human preference data is one pervasive cause of mode collapse. This bias sharpens the probability distribution towards a few stereotypical completions. When many high-quality completions are possible (e.g., in joke generation), this sharpening becomes a tie-breaker, resulting in mode collapse.

3.1 TYPICALITY BIAS IN PREFERENCE DATA: COGNITIVE & EMPIRICAL EVIDENCE

Typicality Bias Hypothesis. Cognitive psychology shows that people prefer text that is *familiar*, *fluent*, and *predictable*. This preference is rooted in various principles. For instance, the *mere-exposure effect* (Zajonc, 1968; Bornstein, 1989) and *availability heuristic* (Tversky & Kahneman, 1973) imply that frequent or easily recalled content feels more likely and is liked more. *Processing fluency* (Alter & Oppenheimer, 2009; Reber et al., 2004) suggests that easy-to-process content is automatically perceived as more truthful and higher quality. Moreover, *schema congruity* theory (Mandler, 2014; Meyers-Levy & Tybout, 1989) predicts that information that aligns with existing mental models will be accepted with less critical thought. We therefore hypothesize that these cognitive tendencies lead to a *typicality bias* in preference data, in which annotators systematically favor conventional text.

Modeling Rewards with Typicality Bias. To capture this hypothesized bias, we model the reward function, which reflects human preferences, as a combination of *true task utility* and *typicality bias*. For a tractable proxy of typicality bias, we employ the log-likelihood from a pretrained base model, $\log \pi_{\text{ref}}(y \mid x)$: as the base model has been trained to maximize likelihood on massive text corpora, its probability scores inherently capture text typicality. Without loss of generality, we use the Bradley-Terry model common in RLHF (Bradley & Terry, 1952; Christiano et al., 2017; Ouyang et al., 2022) and formulate this combination in reward models in Eq. 1:

$$r(x, y) \;=\; r_{\text{true}}(x, y) \;+\; \alpha \,\log \pi_{\text{ref}}(y \mid x) \;+\; \epsilon(x), \tag{1}$$

where r_{true} is the true task utility, α is the typicality bias weight, and ϵ is a noise term. $\alpha > 0$ means that, *holding the true utility fixed*, higher typicality bias increases the reward.

Verifying Typicality Bias in Preference Data. We test this hypothesis on HELPSTEER (Wang et al., 2023b), a preference dataset which provides per-response ratings for both *correctness* (true task utility) and *overall helpfulness* (the final reward). From the training set, we form $6{,}874$ pairs of responses to the same prompt with the same correctness ratings. We then compute their per-token log-likelihoods under both *Llama 3.1 405B Base* and *GLM 4.5 Base*, the base models used as π_{ref}. Fitting these values to Eq. 1, yields $\hat{\alpha} = 0.57 \pm 0.07$ and 0.65 ± 0.07 with the respective base models (both $p < 10^{-14}$). This provides empirical evidence for a positive α in Eq. 1, i.e., human raters are biased towards responses more typical for the base model, independent of correctness (true task utility). See §E.1 and §E.2 for the verification experiments on more preference datasets.

3.2 HOW TYPICALITY BIAS CAUSES MODE COLLAPSE

Having confirmed typicality bias, we need to show how it leads to mode collapse. The RLHF optimization objective under the Bradley-Terry model is as follows,

$$\max_{\pi} \;\; \mathbb{E}_{x\sim\mathbb{D}, y\sim\pi(\cdot\mid x)}\big[r(x, y) - \;\beta\, \mathrm{KL}\big(\pi(\cdot \mid x) \,\|\, \pi_{\text{ref}}(\cdot \mid x)\big)\big] \;, \tag{2}$$

where $\beta > 0$ is the KL coefficient, π_{ref} is the reference policy (e.g., the base model), and π is the learned policy.

Plugging Eq. 1 into the closed-form solution of Eq. 2 (Rafailov et al., 2024) yields an optimum, sharpened by γ (derivation in §E.3):

$$\pi^*(y \mid x) \;\propto\; \pi_{\text{ref}}(y \mid x)^{\gamma} \;\exp\!\left(\frac{r_{\text{true}}(x, y)}{\beta}\right), \qquad \gamma \;:=\; 1 + \frac{\alpha}{\beta} \;>\; 1 \;\text{ when } \alpha > 0. \tag{3}$$

So any positive typicality bias weight α strictly *sharpens* the distribution of π_{ref}. Leaving all else fixed, larger α (stronger typicality in preference data) increases the strength of this effect.

Table 1: Comparison of different prompting methods, given the same computation budget of N total responses. k is the number of candidates generated per LLM call, specified in the prompt (e.g., $k = 5$ for the joke task). y_i denotes the i-th generated candidate, $\hat{p}_i$ denotes its verbalized probability, and $\pi(\cdot|x)$ represents the LLM's output distribution conditioned on the prompt x. For Multi-Turn and VS-Multi, h_{i-1} denotes the conversation history up to turn $i-1$, and t denotes the t-th turn.

Method	LLM Calls	Candidates	Turns	Prompt Example	Definition
1. Instance-level Prompt					
Direct	N	1	1	"Tell a joke about coffee"	$y_i \sim \pi(y\|x)$
CoT	N	1	1	"Think step-by-step, then tell a joke"	$y_i \sim \pi(y\|x_{\text{CoT}})$
2. List-level Prompt					
Sequence	$\lceil N/k \rceil$	k	1	"Tell 5 jokes about coffee"	$(y_1, ..., y_k) \sim \pi(y_1, ..., y_k\|x_{\text{seq}})$
Multi-Turn	N	1	N	Turn 1: "Tell a joke about coffee" Turn 2+: "Tell another joke about coffee"	$y_i \sim \pi(y\|x_{\text{multi}}, h_{i-1})$
3. Distribution-level Prompt (Ours)					
VS-Standard	$\lceil N/k \rceil$	k	1	"Tell 5 jokes with their probabilities"	$(y_1, \hat{p}_1), ..., (y_k, \hat{p}_k) \sim \pi(\cdot\|x_{\text{VS}})$
VS-CoT	$\lceil N/k \rceil$	k	1	"Think step-by-step, then tell 5 jokes with probabilities"	$(y_1, \hat{p}_1), ..., (y_k, \hat{p}_k) \sim \pi(\cdot\|x_{\text{VS-CoT}})$
VS-Multi	$\lceil N/k \rceil$	k	$\lceil N/k \rceil$	Turn 1: "Tell 5 jokes with probabilities" Turn 2+: "Tell 5 more with probabilities"	$(y_1^{(1)}, \hat{p}_1^{(1)}), ..., (y_k^{(t)}, \hat{p}_k^{(t)}) \sim \pi(\cdot\|x_{\text{VS}}, h_{t-1})$

Further, suppose there exists a subset $\mathcal{S}$ of responses such that for all $y, y' \in \mathcal{S}$[1] we have flat true rewards, $r_{\text{true}}(x, y) = r_{\text{true}}(x, y')$[2]. Then by Eq. 3 the optimum within $\mathcal{S}$ reduces to

$$\pi^*(\cdot \mid x) \;\propto\; \pi_{\text{ref}}(\cdot \mid x)^{\gamma} \quad \text{on } \mathcal{S}, \qquad \gamma > 1.$$

This behaves like temperature scaling. As γ grows very large, we will have $y^* \in \arg\max_y \pi_{\text{ref}}(y \mid x)$ for all $y^* \sim \pi(\cdot|x)$ with $y^* \in \mathcal{S}$. This shows that the probability mass is *compressed* toward typical completions (those already favored by π_{ref}), yielding a form of *mode collapse* on set $\mathcal{S}$. Intuitively this means that, when many answers are tied on true task utility (a common scenario in creative writing, social simulation, etc), typicality bias acts as a tiebreaker that sharpens the output of the aligned model into the *mode* of the base model.

4 Method: Verbalized Sampling

We have shown that for a mode-collapsed model, any response $y^* \in \arg\max_y \pi_{\text{ref}}(y \mid x)$ on $\mathcal{S}$, which suggests the need to study the base model π_{ref}. Empirical studies (West & Potts, 2025; Zhu et al., 2025a) have shown that base models do exhibit diversity. Therefore, we propose *Verbalized Sampling* as a prompting strategy to recover the diversity level of π_{ref}, to bypass mode collapse.

4.1 Different Prompts Collapse to Different Modes

For a mode-collapsed LLM, we find that different prompts x collapse to different modes of π_{ref}. This is how VS can mitigate mode collapse. We categorize prompting strategies into three types and provide their corresponding modes. Detailed assumptions and proof are provided in §E.4.

1. **Instance-level prompt**: This is the most traditional prompt x, requesting one instance (e.g., "Tell me a joke about coffee"). The mode is the mode instance (the mode joke) of the base model.
2. **List-level prompt**: This prompt x requests a list of outputs (e.g., "Tell me k jokes about coffee"), as used in Wang et al. (2023a); Dubois et al. (2023). The mode is a uniform distribution of related items (a uniformly-distributed list of jokes) learned by the base model during pretraining.
3. **Distribution-level prompt (ours)**: We propose this prompt x which requests k outputs with corresponding probabilities (e.g., "Tell k jokes about coffee with their probabilities"), and name it ***Verbalized Sampling (VS)***. The mode is a distribution capable of approximating the distribution of related items learned by the base model during pretraining. Figure 2 and §G.9 show that when an LLM is prompted to generate a distribution of the 50 US states, its verbalized probability distribution aligns with a proxy of the same distribution in a pre-training corpus (RedPajama), where the KL divergence is 0.12 for Claude-4-Sonnet.

[1]For example, we can restrict our analysis to $\mathcal{S}$ with only meaningful responses, because nonsensical or erroneous responses are unlikely to be sampled from a well-trained π^*.

[2]This assumption can be relaxed to approximate flatness. We just need bounds on the deviations of r_{true} between y and y' to claim mode collapse, but the overall argument (and result) is consistent.

In Table 1, we summarize how to implement different prompting methods in practice, under the same computation budget of N total generated responses for a fair comparison. In theory, the number of candidates k in each LLM call could be equal to N; but in practice, we notice that if k is too large, the generation quality degrades, so usually $k < N$ and we will generate N total responses across $\lceil N/k \rceil$ calls. For **(2) List-level prompt**, we test another variant, *multi-turn* (West & Potts, 2025), which elicits N responses across N turns in a conversation. For **(3) Distribution-level prompt**, we propose two variants: ***VS-CoT*** and ***VS-Multi***, to further enhance diversity.

4.2 Experimental Setup

LLMs. Our method is training-free, model-agnostic, and requires no logit access. We test it on a suite of models: (1) closed models like GPT Series (**GPT-4.1-mini**, **GPT-4.1**), Gemini Series (**Gemini-2.5-Flash**, **Gemini-2.5-Pro**) and Claude Series (**Claude-3.7-Sonnet**, **Claude-4-Sonnet**); (2) open ones like **Llama-3.1-70B-Instruct** and **Qwen3-235B-A22B-2507-Instruct-2507**; and (3) reasoning models like **OpenAI o3** and **DeepSeek R1**. See §I.1 for generation hyperparameters.

Tasks. We conduct comprehensive experiments on creative writing (§5), dialogue simulation (§6), open-ended QA (§7), synthetic data generation (§8 and §G.6.2), random number generation (§G.5), along with commonsense reasoning (§G.7) and safety (§G.8) to show that our method maintains factual accuracy and safety.

5 Creative Writing

Following prior work on LLM diversity (Lu et al., 2025a), we first study three creative writing tasks: poem continuation, story generation, and joke writing.

Benchmarks. We evaluate model performance on three benchmarks. For (1) poem continuation and (2) story generation, we follow the text continuation setup in Lu et al. (2025a), and use poems from PoemHunter.com and stories from the BookMIA dataset (Shi et al., 2024) for experiments. For (3) joke writing: we follow Turgeman et al. (2025) and curate 100 thematic prompts from the Reddit r/DadJokes dataset (Reddit, 2023), each structured as "Write me a joke about [topic]" (e.g., "...about an octopus"). To reduce computation costs, we randomly select 100 data points for these three tasks, and apply verbalized sampling to generate $k = 5$ candidates and $N = 30$ total samples for each data point. Detailed prompts are provided in Appendix I.2.

Evaluation. We evaluate all methods on two metrics: *diversity* and *quality*. (1) For diversity, we assess both semantic and lexical levels: (i) For semantic diversity, we follow prior work (Cox et al., 2021; Cann et al., 2023; Lu et al., 2025a; Zhu et al., 2025a) and calculate $1 - \bar{s}$, where $\bar{s}$ is the mean pairwise cosine similarity of response embeddings (generated using OpenAI's `text-embedding-3-small` model). Negative similarities are clipped to 0 to avoid inflating diversity and we present the final score as a percentage, where 100% represents maximum diversity. (ii) For lexical diversity, we follow Shaib et al. (2025) and use ROUGE-L (Lin, 2004), where lower scores indicate greater diversity. (2) To evaluate output quality, we use Claude-3.7-Sonnet as the judge. We score *Poem* and *Story* with the rubrics from Creative Writing v3 (Paech, 2023), and jokes with the Humor grader rubrics from HumorBench (Narad et al., 2025a). See Appendix I.3 for details on evaluation.

5.1 Results

Diversity Score. Figure 3(a)-(c) show the semantic diversity score averaged across models on poem, story, and joke, respectively. Across tasks, VS-Standard consistently and significantly outperforms baseline methods. The variants, VS-CoT and VS-Multi, further improve generation diversity. Detailed results on lexical diversity and individual model families are in Appendix G.1.1.

Diversity vs. Quality. Figure 3(d) shows the diversity-quality trade-off on the poem task. The quality of VS-Standard remains comparable to other methods. Notably, VS-CoT achieves the highest diversity while maintaining a high quality score, pushing the Pareto front of this trade-off (Zhang et al., 2021). This shows that VS can boost diversity without harming quality. See Appendix G.1 for the diversity-quality trade-offs for the story and joke tasks.

Emergent Trend. We observe an emergent trend where larger models benefit more from VS. Figure 3(e) shows the diversity gain over the direct prompting which suffers from mode col-

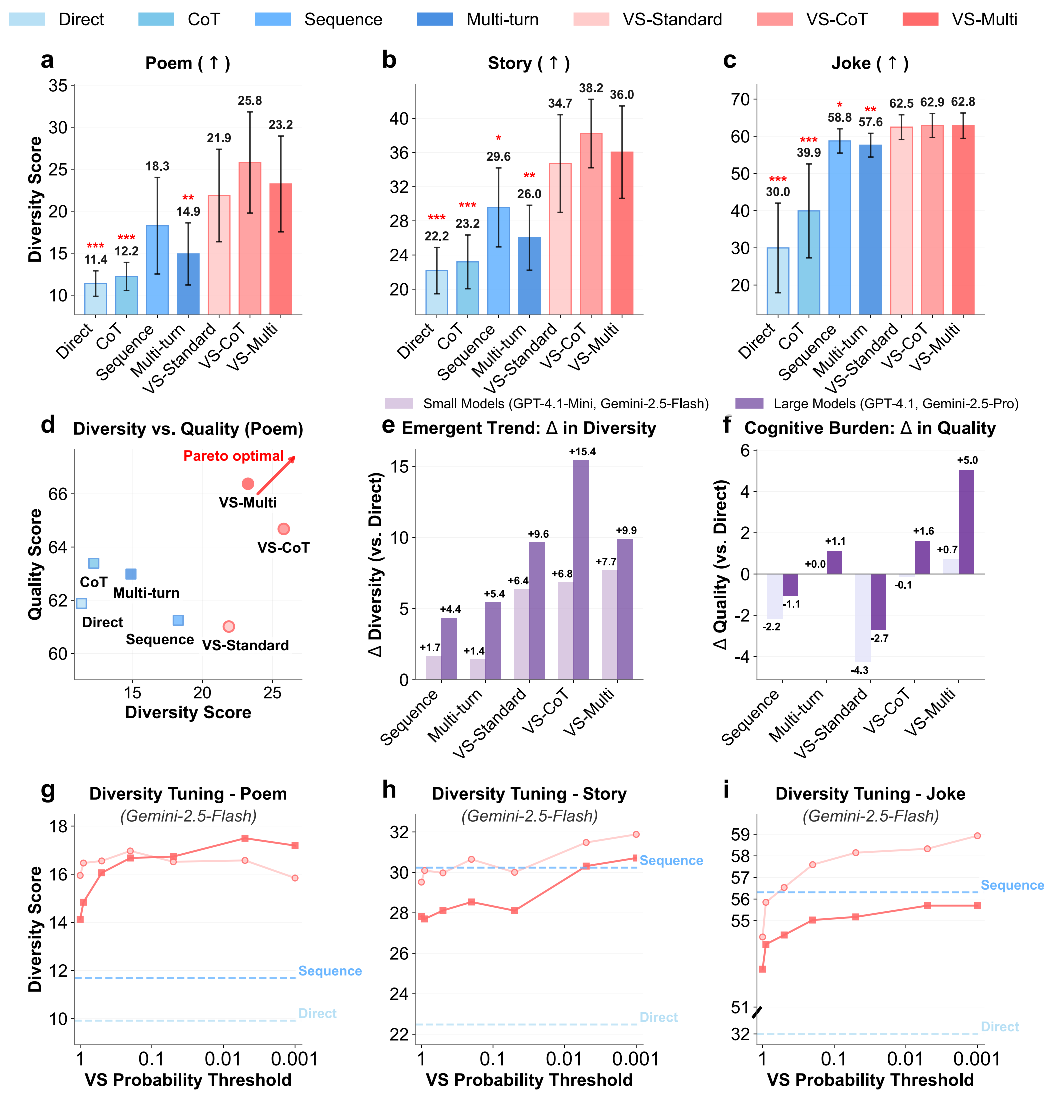

Figure 3: **a-c: Average semantic diversity scores** (%) in poem (**a**), story (**b**) and joke (**c**) across methods and models. Our methods consistently outperform the baselines. We performed a one-tailed t-test between VS-Standard and the baselines (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). **d: Diversity vs. Quality trade-off** for the poem task, where VS-Multi and VS-CoT approach the Pareto front. **e-f: Emergent Trend** where larger models benefit more from VS. We show differences in diversity **(e)** and quality **(f)** over Direct across small (GPT-4.1-Mini, Gemini-2.5-Flash) and large (GPT-4.1, Gemini-2.5-Pro) models. **g-i: Tunable Diversity** shows the diversity tuning results on Gemini-2.5-Flash across tasks. Unlike baseline methods in dashed lines, we can tune the diversity level with VS: as the probability threshold decreases, diversity increases.

lapse. Across all VS variants, larger models (GPT-4.1, Gemini-2.5-Pro) achieve diversity gains 1.5 to 2 times greater than smaller models (GPT-4.1-Mini, Gemini-2.5-Flash).

Cognitive Burden. This scaling trend also extends to quality, as shown in Figure 3(f). While prior work (Hu et al., 2024) found complex prompts create a "cognitive burden" that degrades LLM performance, our findings are nuanced. Methods like Sequence and VS-Standard do cause a drop in quality, but this effect is less severe for larger models. Notably, more intricate variants like VS-CoT and VS-Multi overcome this burden, even improving quality in larger models. This suggests using VS variants may better utilize the capabilities of advanced models, turning complexity into benefits.

Diversity Tuning. Unlike baseline methods, VS allows us to tune the output diversity by adjusting the probability threshold directly in the prompt (e.g., "Generate five responses with probabilities

Table 2: Qualitative comparison for story generation. For the prompt "Without a goodbye", the direct method consistently generates stories about a romantic partner's disappearance. In contrast, our method (VS) demonstrates more creative diversity, generating stories that explore modern communication, metaphorical endings, and cosmic-scale events.

Instruction: "Please write a short story starting with the following prompt: Without a goodbye,"	
Direct	**Our Method (VS)**
Example 1 Without a goodbye, Elara vanished. Her side of the bed was cold, a faint indentation in the pillow the only evidence she'd ever been there. Liam stared at it, a knot tightening in his stomach. No note, no hurried text, just a gaping silence where her vibrant presence used to be. He searched the apartment, a frantic, methodical sweep. Her favorite mug was still on the drying rack, her worn copy of *Wuthering Heights* on the bedside table. Nothing seemed out of place, yet everything felt terribly wrong.	**Example 1** Without a goodbye, the email landed in my inbox. It was from Mark, my best friend since kindergarten... The subject line was simply: "Moving On." ... The message itself was a concise, almost clinical statement that he was relocating to Patagonia... There was no "sorry for the short notice," no "I'll miss you," not even a simple "good luck." ... It was a clean cut, a sudden erasure that left me suspended in disbelief, grappling with the phantom limb of a friendship that had simply, inexplicably, ceased to be.
Example 2 Without a goodbye, Elara vanished. Her side of the bed was cold, the coffee mug she'd used that morning still on the nightstand, a faint lipstick stain on the rim. Liam found the note taped to the fridge, a single, elegant script: "Free at last." He stared at it, the words blurring, then sharpening, then blurring again. Free from what? From him? From the life they'd built, over fifteen years? The house felt too big, too silent.	**Example 2** Without a goodbye, the last star flickered out. It wasn't a violent explosion, no supernova spectacle, just a gentle fade, like a candle finally exhausting its wick. The universe, which had always hummed with the silent symphony of distant suns, suddenly became an echoing void. I was the last observer, adrift in my tiny, self-sustaining craft, my instruments registering only an infinite, inky blackness.
Example 3 Without a goodbye, Elara vanished. Her side of the bed was cold, the coffee machine untouched. A half-read book lay open on the nightstand, a bookmark tucked precisely between pages 107 and 108; a detail that always irritated Liam. She was meticulous, almost to a fault. But this: this was not meticulous. This was an absence, stark and sudden.He checked her usual haunts: the community garden where she nurtured prize-winning tomatoes, the dusty antique shop where she hunted for forgotten treasures, the library where she volunteered on Tuesdays.	**Example 3** Without a goodbye, the music simply stopped. Not a fade-out, not a final crashing cymbal, but an abrupt, jarring silence that left the dancers frozen mid-step. The DJ, a wizened man named Frankie who had spun records at 'The Blue Note' for forty years, had simply collapsed behind his turntables. One moment, the room throbbed with a soulful blues riff; the next, only the murmur of shocked whispers filled the air.

below {threshold}"), without altering decoding parameters. As shown in Figure 3(g-i), diversity increases as the probability threshold decreases. See Appendix H.4 for more detailed results.

Qualitative examples. Figure 2 and Table 2 show qualitative examples on the story task. VS produces more creative stories than direct prompting. See more qualitative examples in §F.

5.2 Human Study on Diversity

To complement our automatic diversity scores, we conducted a human evaluation on Prolific. Following past work, we provided task-specific diversity definitions (plot, style and setup-punchline, respectively). For each task, 30 annotators rated the diversity of 90 output pairs from three prompting methods (Direct, Sequence, VS-Standard) across ten curated topics. Each pair was rated on a four-point Likert scale adopted from Chen et al. (2022): Very Similar, Somewhat Similar, Somewhat Dissimilar, or Very Dissimilar. Inter-annotator agreement was moder-

Table 3: Human-rated diversity (1 = Very Similar, 4 = Very Dissimilar) for poem, story, and joke tasks under Direct, Sequence, and VS-Standard.

Task	Direct	Sequence	VS-Standard
Poem	1.90	2.07	**2.39**
Story	2.74	2.76	**3.06**
Joke	1.83	2.93	**3.01**

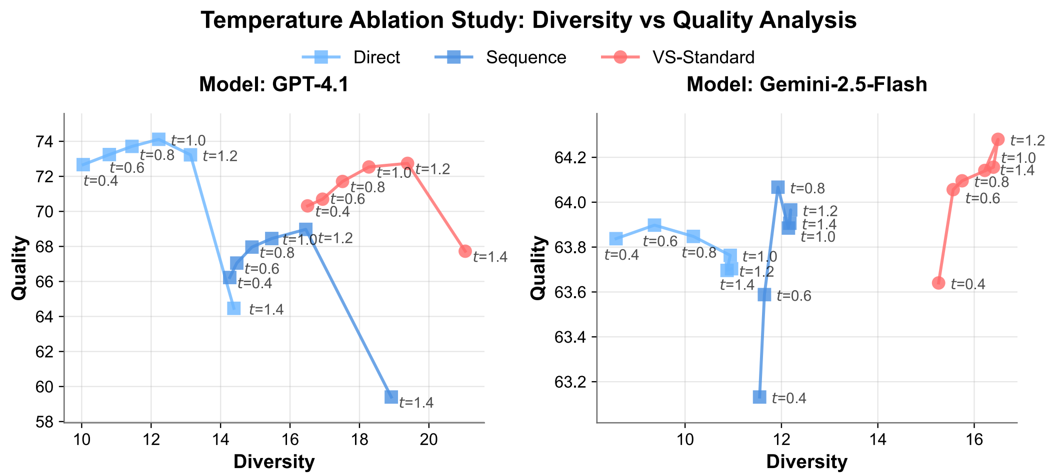

Figure 5: **Ablation study on temperature for poem generation across GPT-4.1 and Gemini-2.5-Flash models.** We set $k = 5$ across experiments. Each plot shows the diversity-quality trade-off for three methods (Direct, Sequence, VS-Standard) at different temperature values (t). VS-Standard can be combined with temperature to further improve the trade-off, consistently outperforming baselines across both models.

ate for poems (0.54), high for stories (0.87) and jokes (0.86). Table 3 shows that VS achieves higher diversity than the baselines on all tasks. See §G.2 for more details on the human study.

5.3 Ablation Study

In this section, we present two ablation studies on the poem task in detail. First, we ablate various post-training stages (SFT, RLHF, RLVR) and show empirical evidence that post-training causes mode collapse and VS can indeed mitigate it and reduce the loss of diversity compared with other methods. Second, we ablate the temperature and show that VS's performance gains are orthogonal to temperature scaling, allowing the two to be combined to further improve the diversity-quality trade-off.

Ablation on VS across post-training stages We employ the Tulu-3 family (Lambert et al., 2025) , which contains checkpoints for SFT, RLHF and RLVR starting from Llama-3.1-70B-base models (Meta, 2024), for the poem task. Figure 4 shows the results: traditional prompting methods do experience much larger diversity drops (*mode collapse*) as models undergo alignment training, and **VS can mitigate mode collapse and maintain a higher diversity score across different post-training stages** (the diversity still drops after SFT, but SFT is necessary for instruction following capability). Specifically, direct prompting exhibits the most severe mode collapse, with diversity dropping from 20.8% after SFT to just 10.8% after DPO. Other methods like sequence and multi-turn prompting also show decreased diversity. In contrast, VS maintains a stable diversity of around 30% across stages. After the DPO stage, VS outperforms direct prompting by 182.6% and retains about 66.8% of the base model's original diversity. Direct prompting, by comparison, retains only 23.8%. This suggests that VS effectively mitigates the mode collapse induced by alignment training.

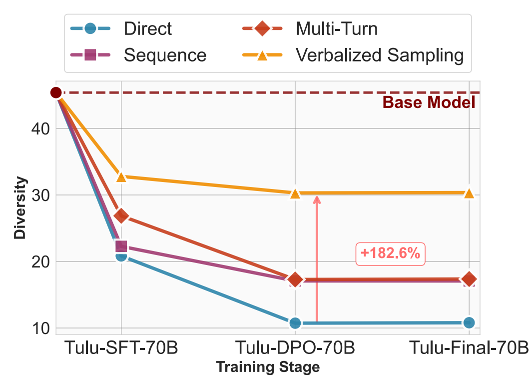

Figure 4: **Diversity scores across post-training stages of Tulu-70B.** "Tulu-Final-70B" is the model after RLVR. The red dashed line indicates the base model's diversity level (45.4%). Baseline prompting methods experience major diversity drops (*mode collapse*) after SFT and DPO, with direct prompting showing the most severe drop. In contrast, VS maintains a higher diversity scores throughout all training stages, demonstrating that it can mitigate *mode collapse*.

Ablation on Temperature. We investigate the effect of sampling temperature on the diversity-quality trade-off. We vary the sampling temperature ($t \in \{0.4, 0.6, 0.8, 1.0, 1.2, 1.4\}$) for three methods (Direct, Sequence, and VS-Standard) across two models (GPT-4.1 and Gemini-2.5-Flash). Figure 5 presents the diversity-quality Pareto front for each method. The results indicate that **VS-Standard can be combined with temperature to further improve the diversity-quality trade-off.** VS consistently achieves a better balance between quality and diversity across both models, pushing forward the Pareto front relative to the direct and sequence baselines.

Ablation on Number of Candidates, Decoding Methods, and Prompt Formats. We also perform comprehensive ablation studies on the poem task on other factors. (1) Appendix H.1 shows that a higher number of candidates, k, leads to greater diversity. (2) In Appendix H.2, we vary the decoding strategies (top-p, and min-p), and show that VS is also orthogonal to these decoding strategies and can be combined with them to further enhance the diversity-quality curve. (3) In Appendix H.3, we test different prompt formats for eliciting distributions (e.g., asking for "probability", "percentage", or "confidence"). While all formats improve diversity, we use the empirically best-performing format in all of our experiments: "probability" for VS-Standard and VS-CoT and "confidence" for VS-Multi. Across all these ablations, VS consistently outperformed the baselines under the same setups.

Takeaway 1: On creative writing tasks, Verbalized Sampling enhances diversity while maintaining quality and allowing tunable diversity. It also better retains diversity through post-training stages and complements different decoding strategies. Notably, larger models benefit more from VS.

6 Dialogue Simulation

Simulating multi-turn dialogues with LLMs is crucial for applications like social simulation (Lin, 2025; Anthis et al., 2025a) and LLM evaluation (Zhou et al., 2024). But existing methods suffer from generic responses and low realism against human dialogues. We therefore test VS on this task.

Benchmark. We use the *PersuasionForGood* task (Wang et al., 2019), which contains 1,017 dialogues where one participant persuades another to donate to the organization, "Save the Children". We choose this dataset as it includes participant personas and a clear, verifiable outcome, the final donation amount, allowing for comparison between the human interactions and our simulation ones. After filtering out dialogues with inconsistent donation amounts, we obtain 939 valid instances, which we partition into 739 for training and 200 for testing.

Experiment Setup. In our experiments, we focus on simulating the persuadee to assess the realism of persuasion outcomes. The model is given a task instruction and a persona to match the human participant. It interacts with a GPT-4.1-based persuader, prompted with the persuader instruction and persona (see Appendix I.2 for prompts). To establish a strong supervised baseline for the simulation, we also fine-tuned Llama-3.1-8B on the persuadee responses in the *PersuasionForGood* training set.

Unlike single-output creativity writing, dialogue simulation is a multi-turn task, so we need to select a response to continue the interaction at each turn. We explore two design choices at each turn: (1) *Number of candidates*: either a model-decided variable or a human-decided constant ($k = 5$); (2) *Response sampling strategy*: probability-weighted (using verbalized probabilities) or random (uniform over candidates). Empirical results show that model-decided random sampling and human-decided probability-weighted sampling best balance the response quality and diversity; so we adopt these two designs in our experiments.

Evaluation. We evaluate our simulation on the *PersuasionForGood* human-human test set across two dimensions: donation amount and linguistic style. (1) For **donation amount alignment**, we compare the human and simulated donation amounts with the (i) Kolmogorov-Smirnov (KS) test (Massey, 1951) for distributional alignment and (ii) L1 distance for per-dialogue alignment. (2) For **linguistic alignment**, we assess three metrics: (i) lexical diversity using Distinct-N (Li et al., 2016), which is the proportion of unique n-grams, (ii) semantic diversity using pairwise embedding-based diversity on persuadee responses within a dialogue, and (iii) readability using the Flesch–Kincaid Grade Level (Flesch, 1948).

6.1 Results

Donation Amount Alignment. Figure 6(a) shows the distribution of donation amounts, with the human ground truth in blue. Across models, VS simulates donation distributions more aligned with

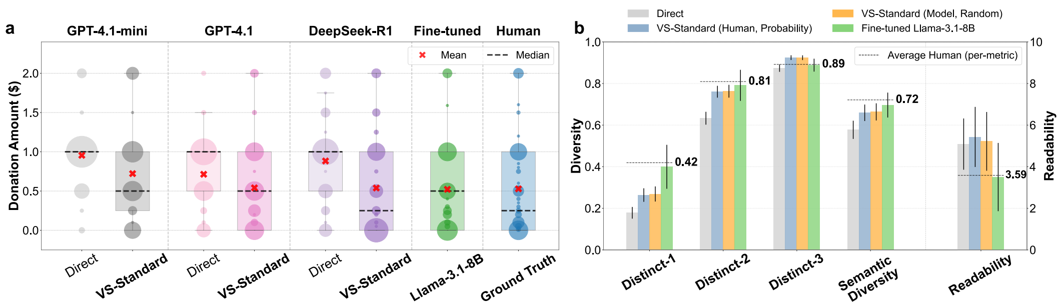

Figure 6: **VS performance in Persuasive Dialogue Simulation. (a) Donation Amount Distributions** simulated by small, large, and reasoning models with direct and VS, compared against fine-tuned model (green) and human (blue). We see that VS simulates donation distributions more similar to human, especially for the larger and reasoning-focused models. **(b) Linguistic Alignment** on Distinct-1/2/3, semantic diversity, and readability. Black dashed lines denote human levels; closer values indicate better stylistic match. VS achieves higher diversity than the direct prompting, approaching human levels. But the readability score remains higher, suggesting room for improvement.

human behaviors than direct prompting. We also observe an *emergent trend* that larger models (e.g., GPT-4.1 vs. GPT-4.1-mini) and reasoning-focused models like DeepSeek-R1 benefit more from VS. Notably, GPT-4.1 with VS matches a fine-tuned Llama-3.1-8B persuadee simulator, and DeepSeek-R1 even surpasses it in simulating the median donation amount. The qualitative example in Figure 1 shows that VS can generate human-like behaviors, such as resistance and changes of mind (see Table 17). We did not evaluate other VS variants due to high simulation costs. Quantitative results on KS tests and L1 distance are provided in Table 23.

Linguistic Alignment. Figure 6(b) shows the results. On the diversity side, VS with different settings (model-decided random sampling and human-decided weighted sampling) outperforms direct prompting on Distinct-1/2/3 and semantic diversity, approaching the fine-tuned model's performance and the human distribution. Qualitative analysis shows that VS simulates more substantive responses than direct prompting (see Table 16 and Table 17). On the readability side, VS still simulates more complex responses than fine-tuned models and humans, suggesting room for improvement. Full results are provided in Table 24.

Takeaway 2: VS helps models better simulate multi-turn dialogues, leading to more diverse conversations and donation distributions that are closer to actual human donation behavior.

7 Open-Ended QA

For real-world tasks like survey simulation, generating a broad and realistic range of answers is crucial. Building on our finding that VS improves diversity, this section evaluates its effectiveness in producing such distributions for open-ended questions with multiple valid answers.

Benchmark. We adapt from the *CoverageQA* (Wong et al., 2024) benchmark, which contains simple QA questions with a wide range of valid answers (e.g., "Name a US state"). Our evaluation uses 40 questions (10 original, 30 new ones created in the same style), each with at least 20 ground-truth answers requiring no reasoning or external knowledge. For each question, we sample $N = 100$ responses per method by generating $k = 20$ candidates per LLM call, capturing both within-call and across-call diversity. Full prompts are in Appendix Appendix I.2.

Evaluation. We evaluate the performance using three metrics: (1) **KL divergence**, the deviation of the model's answer distribution from a realistic reference distribution estimated from the RedPajama (Computer, 2023) pretraining corpus. Lower values indicate better alignment. Note that here we focus on the generated answers rather than the verbalized probabilities, so we calculate the answer distribution from the frequency of each unique answer, not from the verbalized probability distribution like in Figure 2. (2) **Coverage-N**, the fraction of unique ground-truth answers generated in N samples; higher values indicate broader coverage. (3) **Precision**, the proportion of correct answers among all samples; it measures if the increased diversity comes at the expense of correctness.

Results. As shown in Figure 7, our methods outperform all baselines. VS-Standard significantly lowers KL divergence and improves coverage. VS-Multi achieves the best overall tradeoff, yielding

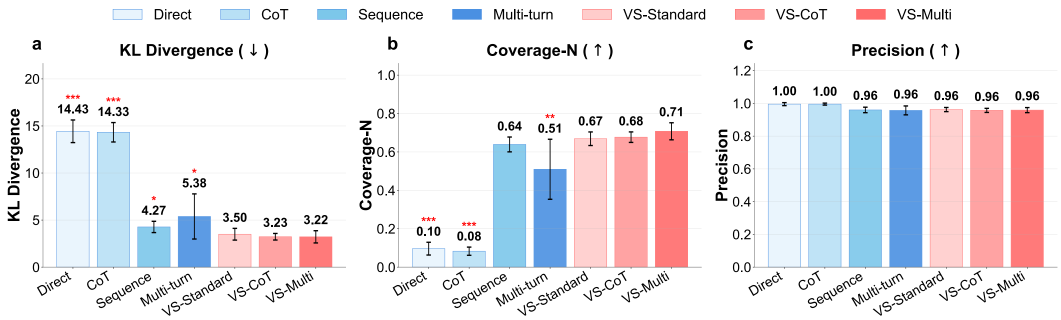

Figure 7: Results on the **Open-Ended QA** task averaged across models. We perform one-tailed t-test between VS-Standard and baselines ($^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$). **(a)** shows the average KL divergence between the response distribution and the corresponding pretraining distribution. VS achieves lower KL divergence compared to baseline methods, indicating closer alignment with the pretraining distribution. **(b)** shows the average Coverage-N across all models. This means VS can generate a broader range of correct answers than the baselines. **(c)** shows the average precision across all models. VS methods maintain answer quality comparable to baseline approaches.

the lowest KL divergence and the highest coverage. Crucially, these gains do not compromise answer quality, as precision remains near 1.0 across all methods. Detailed results are available in Table 25.

Takeaway 3: VS improves alignment with the pretraining distribution and increases answer coverage without compromising answer quality in open-ended QA with multiple valid answers.

8 Synthetic Data Generation

Recent research has shown that the diversity of synthetic data plays an important role in improving downstream model performance (Chen et al., 2024a; Zhu et al., 2025a). So we further evaluate VS on synthetic data generation, including incorrect synthetic data in § G.6.2.

Synthetic Data Generation Setup. We prompt two models, GPT-4.1 and Gemini-2.5-flash, with different prompting methods to generate $N = 1,000$ synthetic competition math questions, with $k = 5$ in each call. We use a small k to ensure the generation quality as it is a complex task. See Appendix I.2 for the prompts. Then we use Qwen3-32B to generate their corresponding reasoning trajectory and answers, as the model is proficient on math benchmarks and capable of producing reliable reasoning traces. See §G.6.1 for more implementation detail.

Fine-tuning on Synthetic Data. With this 1K synthetic dataset, we follow the SFT setting in LIMO (Ye et al., 2025), an effective method to improve reasoning performance with small dataset size, and finetune the following models on this 1K dataset: Qwen2.5-7B, Qwen3-1.7B-Base, and Qwen3-4B-Base (Qwen, 2025a;b).

Benchmarks and Evaluation We evaluate the fine-tuned models' downstream task performance on three widely used math benchmark datasets: MATH500 (Hendrycks et al., 2021), OlympiadBench (He et al., 2024), and Minerva Math (Lewkowycz et al., 2022), which cover a wide range of topics, including algebra, geometry, and competitive mathematics. We use `math_verify`[3] for the evaluation.

Results. Table 4 shows the average accuracy across the three datasets. VS and its variants improve the downstream performance on math tasks across the board, with VS-multi achieving the strongest average accuracy of 37.5%. In contrast, using direct prompting may even hurt the performance due to mode collapse. This suggests that it is a promising direction to apply VS for synthetic data generation to enhance downstream task performance. See Table 27, 28, and 29 in §G.6.1 for the results on individual datasets.

Takeaway 4: VS generates more diverse synthetic data, improving downstream performance on math tasks. This work highlights the capability of LLMs to generate diverse synthetic data, pointing toward a promising paradigm for training more capable models.

[3] https://github.com/huggingface/Math-Verify.

Table 4: Downstream accuracy averaged across MATH500, OlympiadBench and Minerva Math. "Gen Models" show the models used to generate the 1K synthetic questions. "SFT Models" are the ones used to finetune on the 1K synthetic data. VS and its variants improve the downstream task performance.

Gen Model	**GPT-4.1**			**Gemini-2.5-Flash**			
SFT Model	Qwen2.5-7B	Q3-1.7B-Base	Q3-4B-Base	Qwen2.5-7B	Q3-1.7B-Base	Q3-4B-Base	**Average**
Baseline	27.2	30.5	40.7	27.2	30.5	40.7	32.8
Direct	26.1	31.4	34.5	24.9	29.5	36.9	30.6
CoT	30.1	32.5	39.4	27.6	32.1	40.5	33.7
Sequence	30.5	31.0	42.1	28.2	31.7	42.5	34.3
Multi-Turn	29.9	31.9	41.3	27.1	32.2	37.1	33.2
Our Methods							
VS-Standard	32.7	33.6	45.5	28.6	33.3	42.8	36.1
VS-CoT	33.4	33.7	**45.9**	29.4	**35.8**	43.4	36.9
VS-Multi	**34.8**	**34.9**	45.0	**31.7**	34.8	**43.6**	**37.5**

9 Conclusion

This work reveals that mode collapse in aligned LLMs stems from a fundamental property of human preference data: *typicality bias*, the cognitive tendency of human annotators to prefer conventional responses. We formalize this bias theoretically and validate it empirically across multiple preference datasets, confirming its pervasiveness. Grounded in our theoretical understanding, we propose Verbalized Sampling (VS), a simple but principled prompting method that mitigates mode collapse. VS instructs the model to generate a probability distribution over candidate responses, thereby restoring the diverse distribution learned during pretraining. Extensive experiments show that VS significantly enhances performance across tasks (creative writing, dialogue simulation, open-ended QA, synthetic data generation) without compromising factual accuracy or safety. We also identified an emergent trend where stronger models benefit more from VS, suggesting that our method effectively unlocks LLMs' inherent creative potential. This work provides both a novel data-level lens to understand the limitations of various alignment methods and a practical, lightweight solution to overcome mode collapse, paving the way for more creative applications with LLMs.