G2: Guided Generation for Enhanced Output Diversity in LLMs

Zhiwen Ruan[1], **Yixia Li**[1], **Yefeng Liu**[4], **Yun Chen**[3], **Weihua Luo**[4]
Peng Li[2], **Yang Liu**[2], **Guanhua Chen**[1*]
[1]Southern University of Science and Technology, [2]Tsinghua University
[3]Shanghai University of Finance and Economics, [4]Alibaba International Digital Commerce

Abstract

Large Language Models (LLMs) have demonstrated exceptional performance across diverse natural language processing tasks. However, these models exhibit a critical limitation in output diversity, often generating highly similar content across multiple attempts. This limitation significantly affects tasks requiring diverse outputs, from creative writing to reasoning. Existing solutions, like temperature scaling, enhance diversity by modifying probability distributions but compromise output quality. We propose **Guide-to-Generation** (G2), a training-free plug-and-play method that enhances output diversity while preserving generation quality. G2 employs a base generator alongside dual Guides, which guide the generation process through decoding-based interventions to encourage more diverse outputs conditioned on the original query. Comprehensive experiments demonstrate that G2 effectively improves output diversity while maintaining an optimal balance between diversity and quality.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in a wide range of natural language generation tasks (OpenAI et al., 2024; DeepSeek-AI et al., 2024). Despite their fluency and coherence, LLMs frequently generate generic, repetitive, or overly conservative outputs, even when explicitly instructed to produce random or diverse outputs (Zhang et al., 2024b; Lanchantin et al., 2025). This limitation hinders their performance in tasks that demand diverse and informative responses, such as open-domain question answering, instruction following, and test-time scaling for reasoning (Zhang et al., 2024a).

Existing efforts to enhance generation diversity can be broadly categorized into training-time and decoding-time approaches. Training-based techniques modify the loss functions during supervised fine-tuning or reinforcement learning to encourage diverse outputs (Zhang et al., 2024b; Li et al., 2025; Lanchantin et al., 2025), but they often require additional training overhead and are customized for specific tasks, making them inefficient and inflexible. Decoding-based approaches offer greater adaptability but face a fundamental **diversity–quality trade-off**. Sampling-based methods like temperature scaling increase variability by flattening the output distribution (Peeperkorn et al., 2024; Renze, 2024; Zhu et al., 2024), yet they fail to leverage previous responses and may only yield marginal diversity gains. Prompt-based methods enhance diversity by conditioning on prior generations, but repeated prompting often leads to semantic drift, undermining output quality (Zhang et al., 2025).

To address these limitations, we propose Guide-to-Generation (G2), a training-free plug-and-play decoding strategy that enhances output diversity while preserving generation quality. G2 consists of three coordinated modules operating within the same LLM: a **base generator** focused on maintaining response quality, a **Diversity Guide** that encourages novel outputs, and a **Dedupe Guide** that suppresses repetition. All modules share the same backbone and are differentiated by distinct prompting templates tailored to their specific roles.

To further promote diversity, we introduce a **center selection strategy (CSS)** that selects a small, semantically representative subset of prior generations to condition the guiding modules. Rather than using all previous outputs, which may contain redundant or overlapping content and thus dilute the guidance signal, CSS ensures that the diversity and deduplication prompts are anchored in distinct and non-redundant semantic cues. This targeted conditioning helps steer the generation away from previously explored regions of the output space, fostering greater novelty in subsequent responses.

Additionally, to avoid degradation in output quality due to overly aggressive guidance, G2 employs

* Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 14128–14146
November 4-9, 2025

an **entropy-based selective intervention mechanism**. When the model exhibits high confidence in its token predictions, no intervention is applied; when uncertainty is high, guidance signals are selectively introduced. This mechanism allows G2 to intervene precisely where needed, enhancing diversity while preserving fluency and coherence.

Experimental results across creative and subjective generation, instruction-following, translation, summarization, and math tasks demonstrate that G2 significantly improves output diversity while maintaining high response quality, positioning it as a promising technique for enhancing the diversity of LLM-generated content.[1]

2 Related Work

2.1 Diverse Text Generation

Diversity is essential in text generation for improving best-of-N performance (Stiennon et al., 2020), enabling synthetic dataset construction (Raventos et al., 2023), and generating alternatives for unsatisfactory initial outputs (Zhang et al., 2024a; Garces Arias et al., 2024).

Existing methods for enhancing diversity fall into two categories: training-based and training-free. Empirical analyses have demonstrated that fine-tuning with standard cross-entropy loss tends to suppress generation diversity (O'Mahony et al., 2024; Kim et al., 2025). To mitigate this, recent works propose modifying the SFT objective to encourage more varied outputs (Li et al., 2025; Zhang et al., 2024b). Other work improves diversity by refining preference optimization objectives; for example, DivPO (Lanchantin et al., 2025) extends DPO to encourage more diverse generations.

Training-free approaches typically rely on decoding strategies such as temperature sampling (Renze, 2024; Zhu et al., 2024), with some variants applying token-level temperature adjustments. For example, Entropy-Driven Temperature (EDT)(Zhang et al., 2024a) dynamically adjusts temperature based on model entropy, while KLD-based methods(Chang et al., 2023) use KL divergence between two models to guide temperature tuning. Although these techniques can increase diversity, they often ignore prior outputs and may degrade quality when overly high temperatures are applied. Prompt-based methods attempt to enhance diversity by conditioning the model on its previous generations. While effective to some extent, repeated prompting can introduce semantic drift, leading to outputs that are less relevant or coherent relative to the original query (Zhang et al., 2025).

2.2 Contrastive Decoding

Contrastive Decoding (CD) enhances generation quality by leveraging likelihood differences between a strong and a weak language model (Li et al., 2023). Recent CD methods fall into two main types: multi-model approaches (Liu et al., 2024a; Zhou et al., 2024; Wu et al., 2024; Manevich and Tsarfaty, 2024) and single-model techniques using prompt variation (Shi et al., 2024; Pei et al., 2023; Sennrich et al., 2024; Leng et al., 2024). While multi-model setups offer stronger contrasts, single-model methods are more efficient and avoid vocabulary mismatches. For instance, Context-Aware Decoding (Shi et al., 2024) uses contrast between context-aware and context-free generations to reduce hallucinations, and Integrative Decoding (Cheng et al., 2025) uses prior outputs to enhance consistency. Our method introduces a novel use of contrastive signals between the original prompt and tailored prompts (diversifying and deduplication) to achieve both diversity and quality.

3 Methodology

Despite generating high-quality outputs, LLMs often struggle with response diversity, even when explicitly prompted for it (Zhang et al., 2024b). To address this challenge, we propose **G2**, a novel decoding strategy to dynamically guide the generation process towards greater diversity. The overall Algorithm 1 is provided in the Appendix A.

3.1 Overview

Let M denote the language model. For a given query Q and prompt $\mathcal{P}$, our goal is to generate N diverse answers $\{A_1, \ldots, A_N\}$. When generating the i-th answer, A_i, our objective is twofold: to ensure its accuracy in addressing the query Q and its distinctiveness from the set of previously generated answers, $A_{<i} = \{A_1, A_2, \ldots, A_{i-1}\}$.

As shown in Figure 1, the base generator, using model M, produces a logit vector $\mathbf{z}_t \in \mathbb{R}^{|V|}$ for the token at step t, where $|V|$ is the vocabulary size:

$$\mathbf{z}_t = M[\mathcal{P}, Q](\mathbf{x}_{<t}) \tag{1}$$

The base generator is conditioned only on the initial prompt $\mathcal{P}$ and the query Q, and does not observe

[1] Our code is publicly available at https://github.com/sustech-nlp/emnlp25-g2.

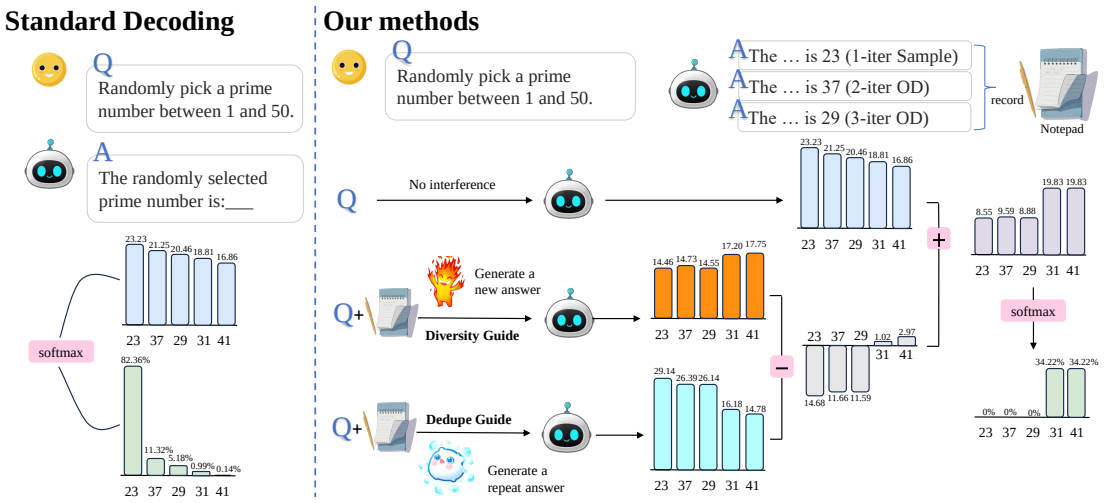


Figure 1: Comparison between standard decoding and our method G2. Standard decoding often produces repetitive outputs, with certain tokens dominating due to peaked softmax distributions. G2 leverages Diversity Guide and Dedupe Guide to encourage diverse and novel generations. See Algorithm 1 in the Appendix for details.

prior responses $A_{<i}$. This ensures that its generation remains focused on the query and unbiased by past outputs during this initial phase.

To steer the generation away from prior responses $A_{<i}$ and encourage novelty, we introduce two Guide components. Leveraging the strong instruction-following capabilities of LLMs, the base generator and these Guides are implemented using the same language model M but are provided with distinct instructional prompts and access to $A_{<i}$. The Diversity Guide is prompted by $\mathcal{P}^+$ to generate content that diverges from $A_{<i}$, while the *Dedupe Guide* is prompted by $\mathcal{P}^-$ to generate content that is similar to $A_{<i}$. Detailed prompt templates for $\mathcal{P}^+$ and $\mathcal{P}^-$ are provided in Appendix D. Their respective logits at token t are computed as:

$$\mathbf{z}_t^+ = M[\mathcal{P}^+, Q, A_{<i}](\mathbf{x}_{<t}) \quad (2)$$

$$\mathbf{z}_t^- = M[\mathcal{P}^-, Q, A_{<i}](\mathbf{x}_{<t}) \quad (3)$$

To combine the guidance from both Guides, we integrate their logits with the generator's output to modulate the final token probability distribution:

$$\mathbb{P}(X_t|\mathbf{x}_{<t}) = \text{softmax}(\mathbf{z}_t + \alpha_t(\mathbf{z}_t^+ - \mathbf{z}_t^-)) \quad (4)$$

where $\mathbb{P}(X_t|\mathbf{x}_{<t})$ is the probability distribution for the next token X_t given the preceding tokens $\mathbf{x}_{<t}$. The core directional guidance for diversification is provided by the contrastive signal $(\mathbf{z}_t^+ - \mathbf{z}_t^-)$. The Diversity Guide $(\mathbf{z}_t^+)$ is prompted to assign higher logits to tokens that introduce novelty with respect to previous answers $A_{<i}$, and conversely, lower logits to tokens likely to cause repetition. The Dedupe Guide $(\mathbf{z}_t^-)$, on the other hand, assigns higher logits to tokens that would closely mirror content in $A_{<i}$ (e.g., tokens '23', '37', '29' as illustrated in Figure 1). Therefore, the difference $(\mathbf{z}_t^+ - \mathbf{z}_t^-)$ serves to amplify the logits of novel tokens while suppressing those of repetitive ones, effectively guiding the model towards generating a distinct output.α_t is a dynamic scaling factor that controls the strength of this Guide's influence on the final distribution.

3.2 Selective Intervention Strategy

While the Guide mechanism provides signals for diversity, determining the appropriate level and timing of intervention is crucial for maintaining quality. Overly aggressive or misapplied guidance can degrade output quality. Therefore, G2 applies intervention selectively at each token generation step based on model uncertainty. We introduce an entropy-based gating strategy to compute the dynamic weighting factor α_t (from Equation 4) for each token X_t.

We quantify model uncertainty using the entropy $H_t = H(\mathbb{P}(X_t|\mathbf{x}_{<t}))$ of its predictive distribution. Intervention is then gated by an entropy threshold β and applied with a fixed strength θ. If $H_t < \beta$, the model is considered confident, and no intervention occurs ($\alpha_t = 0$), preserving high-confidence predictions. If $H_t \geq \beta$, indicating sufficient uncertainty, the Guide mechanism is activated by setting $\alpha_t = \theta$. Formally, α_t is defined as:

$$\alpha_t = \begin{cases} \theta & \text{if } H(\mathbb{P}(X_t|\mathbf{x}_{<t})) \geq \beta \\ 0 & \text{if } H(\mathbb{P}(X_t|\mathbf{x}_{<t})) < \beta \end{cases}$$

This strategy offers advantages over raw entropy

or fixed-weight interventions by better preserving quality. Specifically, it avoids intervening on confident (low-entropy) predictions, and the cap θ prevents degradation from excessive interventions, particularly on high-entropy tokens. The ablation study (Section 5.1, Figure 3) confirms that these alternative strategies result in quality drops.

The threshold β (fixed at 0.1 in all our experiments) and intervention strength θ (e.g., 0.3 or 0.5, selected based on validation experiments detailed in Appendix B.1; $\theta = 0$ disables all intervention) help manage the diversity-quality trade-off. This selective application strategically promotes diversity primarily at points of model uncertainty, thereby balancing novelty with coherence and quality.

3.3 Representative Prior Response Sampling

When generating the n-th response, A_n, conditioning the Guides on all $n-1$ prior responses $(A_1, A_2, \ldots, A_{n-1})$ can result in excessively long prompts. Such lengthy contexts may degrade model performance and are often unnecessarily verbose due to semantic redundancies frequently present within the set of prior responses, $A_{<n}$. Thus, selecting a diverse and representative subset of these prior responses is a more effective strategy.

To facilitate this selection, we first derive semantic embeddings for each prior response $A_j \in A_{<n}$. To avoid reliance on external models, we utilize the base LLM itself for this task. Specifically, each response A_j is processed using the instructional prompt: "This sentence: $\{A_j\}$ means in one word:", and its sentence embedding is extracted from a designated hidden state of the LLM's output. The confluence of a typically small set of prior responses and their high-dimensional embeddings often renders conventional clustering methods suboptimal for selecting a genuinely diverse and representative subset of exemplars.

We therefore employ a greedy iterative method, termed the Center Selection Algorithm, to curate this representative subset. The algorithm initializes by selecting an initial response from $A_{<n}$ (e.g., the most recent or a random one). Subsequently, it iteratively adds the response from the remaining unselected pool that exhibits the maximum dissimilarity (e.g., maximizing the minimum cosine dissimilarity based on their embeddings) to any response already included in the representative set. This process continues until a predefined number of representative responses are selected. This condensed subset then serves as the context of prior outputs for the Guide prompts.

4 Experiments

This section presents comprehensive experiments evaluating G2 across diverse NLP tasks (creative generation, instruction-following, translation, and summarization) for a multifaceted assessment.

4.1 Experimental Setup

This section outlines the evaluation protocols, including diversity metrics, baseline methods, and implementation details employed across our experiments. Task-specific quality metrics are detailed within their respective benchmark discussions.

Diversity Evaluation We assess generation diversity using a suite of established metrics, ensuring a comprehensive evaluation from both lexical and semantic perspectives. Higher values consistently indicate greater diversity.

• **Div-BLEU**: Calculated as 1 − Self-BLEU (Zhu et al., 2018).

• **EAD (Expectation-Adjusted Distinct N-grams)**: Counts distinct 1- to 5-grams, adjusted to mitigate bias from shorter outputs (Li et al., 2016; Liu et al., 2022).

• **Sent-BERT**: Measures semantic diversity as 1 − average cosine similarity between Sentence-BERT embeddings of the responses (Kirk et al., 2024).

These three metrics are aggregated into a composite diversity score, diversity (Div), which gives equal (50%) weight to statistical diversity (EAD and Div-BLEU combined) and semantic diversity (Sent-BERT). The formulation is:

$$Div = \frac{\text{EAD} + \text{Div-BLEU}}{4} + \frac{\text{Sent-BERT}}{2} \quad (5)$$

Baselines We compare our approach against the following baselines. All methods, including ours, utilize Llama-3-8B-Instruct (AI@Meta, 2024) as the backbone LLM to ensure fair comparison.

• **Fixed Temperature Sampling**: Standard sampling with a fixed temperature T, Top-K (50), and Top-P (1.0). T is varied for trade-off analysis, where T $\in \{1.0, 1, 1, 1.2, 1.3, 1, 4, 1.5\}$.

• **Top-P Sampling**: Employs a high temperature (T=1.5) to encourage diversity, with the Top-P (P $\in \{0.8, 0.85, 0.9, 0.95\}$) threshold varied to modulate quality.

• **Top-K Sampling**: Similar to Top-P, it uses a high temperature (T=1.5), varying the Top-K (K $\in \{5, 10, 20, 40\}$) parameter.

Methods	Div-BLEU	EAD	Sent-Bert	Diversity (↑)	Distinct (↑)	Quality (↑)
Claude-3.5 Sonnet	36.27	47.96	19.68	30.90	1.50	8.60
gpt-4o	36.64	50.08	21.53	32.45	1.95	9.01
Llama-3.3-70B-Instruct	32.24	44.66	17.78	28.12	1.70	8.57
Llama3-8B-Instruct	53.59	69.22	29.4	45.40	4.02	7.97
w. Temperature (T=1.3)	62.29	73.61	34.44	51.20 (+5.80)	4.63 (+0.61)	7.69 (-0.28)
w. Top-P (T=1.5, P=0.95)	64.60	75.47	36.44	53.24 (+7.84)	5.17 (+1.15)	7.61 (-0.36)
w. Top-K (T=1.5, K=10)	64.65	75.86	36.17	53.21 (+7.81)	5.08 (+1.06)	7.63 (-0.34)
w. Min-P (T=1.5, P=0.01)	65.24	75.09	35.62	52.89 (+7.49)	5.25 (+1.23)	7.61 (-0.36)
w. EDA (T=1.5, θ=0.1)	60.81	73.93	33.73	50.55 (+5.15)	5.06 (+1.04)	7.61 (-0.36)
w. Diverse Prompt (T=1.0)	57.40	73.90	**40.53**	53.09 (+7.69)	**6.58** (+2.56)	4.73 (-3.24)
w. G2 (θ=0.3)	**64.72**	**78.01**	37.91	**54.64** (+9.24)	5.80 (+1.78)	**7.79** (-0.18)
Qwen2.5-7B-Instruct	56.78	69.42	32.84	47.97	3.60	7.50
w. Temperature (T=1.3)	**66.25**	75.04	34.73	52.69 (+4.72)	4.60 (+1.00)	7.03 (-0.47)
w. Top-P (T=1.5, P=0.95)	64.76	75.72	36.20	53.22 (+5.25)	4.84 (+1.24)	6.99 (-0.51)
w. Top-K (T=1.5, K=10)	64.51	76.80	35.30	52.98 (+5.01)	4.90 (+1.30)	7.20 (-0.30)
w. Min-P (T=1.5, P=0.01)	66.29	75.0	34.54	52.60 (+4.63)	4.82 (+1.22)	6.99 (-0.51)
w. EDA (T=1.3, θ=0.1)	62.82	73.28	34.49	51.27 (+3.30)	4.50 (+0.90)	**7.21** (-0.29)
w. Diverse Prompt (T=1.0)	56.38	74.12	37.78	51.52 (+3.55)	**6.29** (+2.69)	4.03 (-3.47)
w. G2 (θ=0.15)	65.53	**78.79**	**38.04**	**55.10** (+7.13)	5.46 (+1.86)	**7.21** (-0.29)

Table 1: Performance comparison of decoding methods on the NoveltyBench benchmark. While the Diverse Prompt baseline achieves higher Distinct diversity scores, its Quality significantly deteriorates compared to our method (G2). For an expanded comparison across more parameter settings, see Figure 6 in Appendix B.2.

• **Min-P Sampling**: A dynamic truncation method that adjusts the sampling threshold based on the model's confidence (Minh et al., 2025), which uses a high temperature (T=1.5), varying the Min-P $\in \{0.01, 0.03, 0.05, 0.07\}$.

• **Diverse Prompt**: Leverages a prompting strategy (Zhang et al., 2025) that conditions the LLM on previous outputs to encourage varied responses.

• **EDT**: This method (Zhang et al., 2024a) dynamically adjusts temperature based on token-level entropy. It uses an initial temperature T and a fixed adjustment strength $\theta = 0.1$ (consistent with the original work). For WMT'14 and XLSum, we set $T \in \{1.2, 1.4, 1.6, 1.8, 2.0\}$ to allow broader exploration; for other tasks, $T \in \{1.0, 1.3, 1.5\}$.

Details Unless specified otherwise, all methods use Top-K=50 and Top-P=1.0. We generate ten outputs per query for NoveltyBench (following its standard protocol) and five per instance for other benchmarks. For all methods, the first of the multiple outputs is obtained via greedy decoding to establish a consistent quality baseline, while subsequent outputs are generated using the respective method's sampling strategy. For G2, the intervention strength parameter θ is selected from the set $\{0.15, 0.3, 0.5, 0.7\}$, based on validation experiments detailed in Appendix B.1, and the temperature is fixed at 1.0. All experiments are conducted on NVIDIA A100-80G GPU.

4.2 Creative and Subjective Generation Task

4.2.1 Experimental Setup

Benchmark We employ NoveltyBench Curated (Zhang et al., 2025), a benchmark specifically designed with prompts that elicit multiple valid and diverse answers. Its curated dataset spans creative writing, randomness, factual knowledge, and subjective opinion generation, making it ideal for assessing diversity in open-ended tasks.

Diverisity Metric In addition to the common metrics defined in Section 4.1, we report NoveltyBench's specialized **Distinct** metric, which quantifies the number of unique equivalence classes among N generated responses.

Quality Metric Following NoveltyBench, generation quality is assessed using Skywork-Reward-Gemma-2-27B-v0.2 (Liu et al., 2024b) as an automated reward model. We report the average reward score over N responses for each query.

4.2.2 Main Results

Table 1 summarizes the performance of various models and decoding strategies on NoveltyBench. We highlight several key observations:

State-of-the-art proprietary models (Claude-3.5 Sonnet, GPT-4o) and large open-source models

(Llama-3.3-70B-Instruct) achieve high generation quality. However, their output diversity is notably constrained, evidenced by low Distinct scores (1.50, 1.95, 1.70). This underscores the prevalent challenge of diversity in highly capable models, corroborating observations by Lanchantin et al. (2025).

Conventional diversity-enhancement techniques, such as increasing sampling temperature (e.g., T=1.3), can improve diversity and distinct score by 5.8 and 0.61 points, respectively. However, this typically incurs a quality penalty, exemplified by a 0.28 point decrease in the quality score. While Top-K/P and Min-P sampling aim to mitigate this quality degradation at higher temperatures, they often achieve this by tempering the diversity gains. EDT (Zhang et al., 2024a) provides fine-grained, token-level temperature adjustments; yet it fails to leverage previous responses and may only yield marginal diversity gains. These results highlight the ongoing challenge of simultaneously maximizing diversity and quality.

The Diverse Prompt strategy, by conditioning on previous outputs, substantially boosts diversity metrics. However, this comes at a steep cost to output quality (e.g., quality score plummets 3.47 points), rendering it impractical for scenarios demanding high quality. This highlights the difficulty of maintaining relevance and coherence when aggressively prompting for novelty over multiple iterations.

In contrast, **G2** lies near the Pareto frontier, striking an effective trade-off between diversity and quality. On Llama3-8B-Instruct, G2 reaches the highest diversity score and a strong distinct score, while maintaining high quality. Similar patterns are observed on Qwen2.5-7B-Instruct (Team, 2024), where G2 consitently outperforms other methods in balancing novelty with coherence. As shown in Figure 6 (Appendix B.2), G2 reliably occupies the upper-right region of the diversity-quality space, demonstrating its effectiveness in navigating the trade-off between generating varied outputs and preserving fluency and relevance.

4.3 Instruction-Following Task

4.3.1 Experimental Setup

Benchmark We evaluate our methods on two widely used instruction-following benchmarks: AlpacaEval 2.0 (Dubois et al., 2024) and MT-Bench (Zheng et al., 2023). AlpacaEval 2.0 is a single-turn dialogue task where the evaluation metric is the length-controlled win rate (LCWR), which adjusts for the inherent length bias of the judging model. MT-Bench is a two-turn dialogue task, where the evaluation metric is the average score.

AlpcaEval 2.0	Div-BLEU	EAD	Sent-Bert	Diversity (↑)	LCWR (↑)
Llama3-8B-Instruct	53.52	70.13	18.46	40.14	32.01
w. Temperature (T=1.3)	62.02	75.54	20.75	44.77	28.81
w. Top-P (T=1.5, P=0.95)	62.55	75.87	20.98	45.10	27.96
w. Top-K (T=1.5, K=10)	63.71	76.64	20.85	45.51	27.92
w. Min-P (T=1.5, P=0.01)	64.20	76.73	20.84	45.65	27.63
w. EDA (T=1.5, θ=0.1)	65.04	77.07	21.43	46.24	25.64
w. G2 (θ=0.5)	64.10	75.91	24.00	**47.00**	**29.20**
MT-Bench	**Div-BLEU**	**EAD**	**Sent-Bert**	**Diversity (↑)**	**Score (↑)**
Llama3-8B-Instruct	55.47	70.36	22.65	42.78	7.18
w. Temperature (T=1.3)	62.93	74.73	25.38	47.11	**6.96**
w. Top-P (T=1.5, P=0.95)	64.76	76.19	25.86	48.17	6.84
w. Top-K (T=1.5, K=10)	64.21	76.13	25.88	48.03	6.89
w. Min-P (T=1.5, P=0.01)	66.07	77.21	26.16	48.90	6.86
w. EDA (T=1.5, θ=0.1)	66.19	77.60	26.08	48.99	6.75
w. G2 (θ=0.5)	66.55	76.06	27.60	**49.45**	6.92

Table 2: Performance comparison on instruction-following tasks. For an expanded comparison across more parameter settings, see Figure 7 in Appendix B.3

Quality Metrics We use GPT-4o-2024-08-06 (OpenAI et al., 2024) as the judge model and calculate the average score of N responses per query.

4.3.2 Main Results

Table 2 summarizes results for instruction-following tasks, with further visualizations in Figure 7 (Appendix B.3). These findings reinforce G2's consistent advantages.

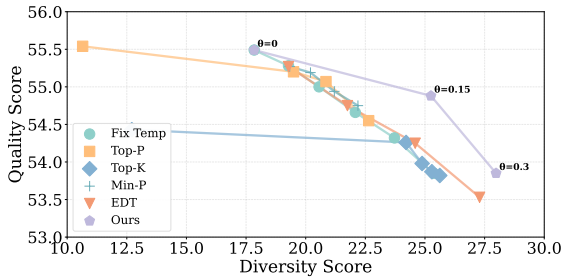
On AlpacaEval 2.0, G2 achieves the highest diversity among the tested methods while maintaining a strong quality score (LCWR). This demonstrates a more effective diversity-quality balance compared to baselines like Top-K sampling or Temperature scaling, which exhibit a more pronounced decline in quality when attempting to reach similar high levels of diversity. Similarly, on MT-Bench, G2 again leads in Diversity while maintaining a competitive quality score. Compared to Temperature scaling (T=1.3), G2's score is only 0.04 lower, yet its diversity is 2.34 points higher. Moreover, both the quality and diversity of G2 are higher than other approaches, such as EDA.

These results are corroborated by visual analyses (Figure 7), where G2's performance curve consistently occupies the desirable top-right region of diversity-quality plots, signifying its superior trade-off characteristics relative to existing baselines.

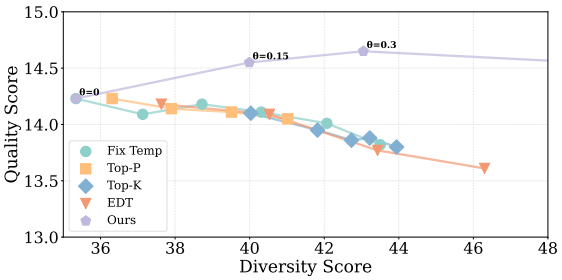
4.4 Translation and Summarization Task

4.4.1 Experiment Setup

Benchmark We evaluate translation on WMT'14 German-English (de→en; 3,003 sentence pairs)



(a) WMT'14 German-English



(b) XLSum

Figure 2: Diversity-quality curves on WMT'14 and XLSum between G2 and other baseline under different settings.

(Bojar et al., 2014) and summarization on 1,000 English instances from XLSum (Hasan et al., 2021).

Metrics Translation quality is assessed by the average of BLEU[2] (Papineni et al., 2002) and COMET[3] (Rei et al., 2020). Summarization quality uses the average of ROUGE-1, ROUGE-2, and ROUGE-L[4] (Lin, 2004). Diversity is evaluated as in Section 4.1.

4.4.2 Main Results

Figure 2 illustrates the diversity-quality relationship for G2 and baselines on WMT'14 de→en and XLSum. Across both benchmarks, G2 consistently operates in the top-right region, signifying a superior trade-off between generation quality and output diversity compared to other evaluated methods.

It is noteworthy that translation and summarization tasks are more constrained than the open-ended tasks like creative generation or instruction-following. These tasks necessitate strict adherence to source content, rendering them less sensitive to standard diversity-enhancement techniques like temperature scaling at moderate settings. Consequently, standard techniques like temperature scaling often need to be set to considerably higher values (e.g., T = 1.5) to yield substantial diversity gains in these contexts. However, such aggressive temperature scaling typically leads to a severe degradation in output quality. This context underscores the effectiveness of G2 in successfully navigating this challenging landscape to improve diversity while preserving the fidelity and quality of the generated translations and summaries.

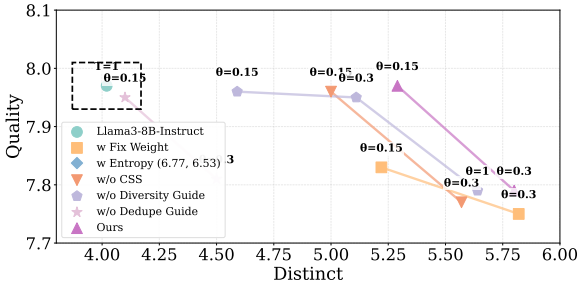


Figure 3: Ablation results of intervention strategy, center selection, and guide components on NoveltyBench.

5 Analyses

5.1 Ablation Study

We conduct ablation studies to evaluate contributions of key components in G2 on NoveltyBench.

Intervention Strategy We compare three strategies: *w Fix Weight*, which intervenes on all tokens with fixed weights ($\theta = 0.15$ and 0.3); *w Entropy*, which uses raw entropy as weights; and G2, which selectively intervenes based on entropy with capped strength. As shown in Figure 3, *w Fix Weight* achieves similar diversity to G2 under the same θ, but G2 yields better quality, especially at $\theta = 0.15$, while intervening on approximately 45% fewer tokens. *w Entropy* improves diversity but significantly reduces quality due to overly large weights on high-entropy tokens, demonstrating the advantage of selective intervention in G2.

Center Selection Algorithm We compare G2 with and without the Center Selection Strategy (CSS), where *w/o CSS* randomly selects past responses instead of using representative ones. As shown in Figure 3, G2 achieves higher diversity under similar quality, indicating that selecting repre-

[2] https://www.nltk.org
[3] https://github.com/Unbabel/COMET
[4] https://pypi.org/project/rouge-score

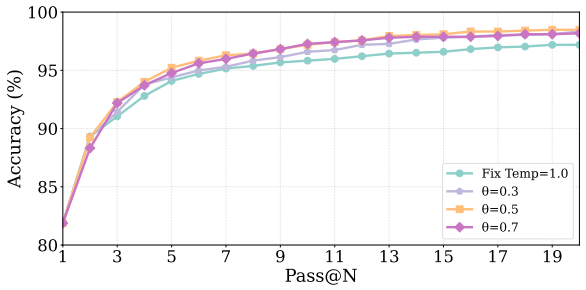


Figure 4: Comparison of Pass@N accuracy on GSM8K between our method and the baseline.

sentative responses via CSS helps guide the model to generate more novel content.

Guide We ablate the two Guide modules: the Diversity Guide and the Dedupe Guide. *w/o Diversity Guide* removes the diversity constraint, while *w/o Dedupe Guide* removes the dedupe constraint. As shown in Figure 3, both variants reduce diversity compared to G2, with a more notable drop when removing Dedupe Guide. We hypothesize that this is because prompting the model to generate repeated tokens (via consistency prompts) is easier than generating novel ones, making it more effective for contrastive decoding to suppress those repetitions and thereby enhance diversity. Overall, using both Guides leads to the best trade-off.

5.2 Enhancing Mathematical Reasoning via Diverse Candidate Generation

For complex reasoning tasks such as mathematical problem-solving, relying on a single generation path may not consistently yield the correct solution. Generating a diverse set of candidate solutions effectively utilizes additional inference-time computation, a technique well-documented to enhance LLM output quality, particularly for reasoning tasks (Welleck et al., 2024; Snell et al., 2025). We evaluate mathematical reasoning capabilities using the GSM8K benchmark (Cobbe et al., 2021) with Llama-3-8B-Instruct. For each query, we sample N outputs and report Pass@N accuracy, which measures whether at least one correct answer appears among the samples. Unlike the BoN approach with a reward model, Pass@N reduces the reliance on the reward model's performance, providing a more direct assessment of the model's ability to generate correct answers.

The hyperparameter θ allows for a nuanced control: $\theta = 0.3$ setting tends to prioritize the quality of individual responses, whereas $\theta = 0.7$ steers the model towards greater diversity among responses.

Methods	Temp	θ	ATLP	Quality	Diversity
Fix Temp	1.0	-	-0.64	55.49	17.84
G2	1.0	0.15	-0.72	54.88	25.25
Fix Temp	1.3	-	-0.74	54.66	22.08
G2	1.0	0.3	-0.83	53.85	27.98
Fix Temp	1.5	-	-0.86	53.85	25.36

Table 3: Comparison of average token log-probability (ATLP), output quality, and diversity on WMT'14.

As shown in Figure 4, our method consistently outperforms the baseline when $N \geq 3$, demonstrating its effectiveness in improving reasoning performance. The high Pass@N performance illustrated in Figure 4 underscores that a strategic balance between the intrinsic quality of individual solutions and the collective diversity across multiple candidates is paramount for maximizing performance on such reasoning tasks.

5.3 Correlation Between Generation Probability and Diversity

To better understand the underlying reason why G2 improves generation diversity, we examine whether it achieves this by deviating from the model's output distribution or by exploring diverse yet likely generation paths that remain aligned with the distribution. We conduct experiments on the WMT'14 German-English translation task (following the setup in Section 4.4) and compare G2 with fixed-temperature sampling.

We use the average token log-probability (ATLP) to measure how closely generated outputs align with the model's predicted distribution. As shown in Table 3, ATLP is positively correlated with output quality across decoding strategies—higher-likelihood generations generally exhibit better quality. Within the same decoding method, a common trade-off is observed: increasing diversity (e.g., through higher temperature) typically leads to lower ATLP and reduced quality. However, this pattern does not always hold across different methods. Notably, G2 with $\theta = 0.15$ yields both higher ATLP and greater diversity than fixed-temperature decoding with $T = 1.3$. These findings indicate that G2 promotes diversity by following varied yet probable generation paths, rather than deviating from the model's distribution.

5.4 Efficiency Analysis

To better understand the runtime characteristics of G2, we benchmarked its **inference latency** relative to common decoding strategies on NoveltyBench.

Method	Prior Responses (N)	Latency (× baseline)
Standard temperature sampling	–	1.00
Top-P / Top-K / Min-P	–	≈ 1.00
EDT	–	≈ 1.05
G2 (sequential)	1 / 2 / 3	2.52 / 2.66 / 2.67
G2 (parallel)	1 / 2 / 3	1.19 / 1.26 / 1.31

Table 4: Relative inference latency of G2 compared to standard decoding methods on NoveltyBench.

Since G2 requires the base generator, the Diversity Guide, and the Dedupe Guide, the additional computation cost is a key factor to evaluate. Table 4 summarizes the latency results, reported relative to standard temperature sampling.

In the **sequential** setting, the base generator, Diversity Guide, and Dedupe Guide are queried as three consecutive forward passes. Even in this case, the latency overhead remains below $3\times$, and is mitigated by G2's selective intervention mechanism, which only applies guidance at high-entropy (uncertain) tokens. This keeps the intervention sparse and avoids unnecessary computation on confident predictions. In the **parallel** setting, the three prompts are batched into a single forward pass, substantially reducing latency. With this configuration, G2 incurs only a modest runtime increase—at most $1.31\times$ relative to baseline decoding, even when conditioned on three prior responses.

Overall, G2 achieves significant gains in diversity with limited runtime overhead. Since it is training-free, model-agnostic, and compatible with optimized inference backends such as vLLM, as well as quantized models, the additional computation is well within practical limits for real-world deployment.

5.5 External Embedding Analysis

In this section, we investigate the impact of replacing the base LLM's embedding space with external alternatives for center selection in G2. Our choice to adopt the LLM's own embeddings was motivated by the desire to keep G2 self-contained and easy to integrate. The embedding extraction module, however, is designed as a plug-and-play component, enabling the use of external embedding models if desired. Although the LLM's embedding space may not always perfectly reflect semantic diversity, G2 is flexible enough to accommodate alternative representations.

To assess this flexibility, we employed the `bert-large-nli-stsb-mean-tokens` model (Reimers and Gurevych, 2019) as an external embedding extractor within the Center Selection strategy. We re-evaluated G2 on NoveltyBench with this external embedding while keeping all other experimental settings unchanged. The results are shown in Table 5.

Configuration	Distinct (↑)	Quality (↑)
G2 (θ = 0.15, LLM embeddings)	5.29	7.97
G2 (θ = 0.15, External embeddings)	5.24	7.93
G2 (θ = 0.3, LLM embeddings)	5.80	7.79
G2 (θ = 0.3, External embeddings)	5.92	7.84

Table 5: Comparison of G2 performance using the LLM's own embeddings versus external embeddings for center selection on NoveltyBench.

The results indicate that G2 remains effective when paired with external embeddings, offering flexibility without compromising performance. In some cases, external embeddings yield slightly better results, suggesting that the center selection mechanism is robust across embedding spaces. Additional analyses are provided in Appendix C.

6 Conclusion

In this paper, we present G2, a plug-and-play method that requires no additional models or training and is designed to enhance the diversity of LLM outputs through the use of Guides. Our experiments demonstrate that G2 effectively improves output diversity while maintaining an optimal balance between diversity and quality. Furthermore, our experiments reveal that G2 can be seamlessly transferred to different tasks. We hope these findings provide a promising avenue for advancing LLM output diversity and encourage further exploration in this field.

Limitations

Although G2 can enhance the output diversity of LLMs through the Guide, a pertinent limitation remains that nuanced control over the generation process via Guide components necessitates underlying models with strong instruction-following and role-playing capabilities. Consequently, the achievable level of control and the finesse of the generated output are dependent on these specific model aptitudes.

Acknowledgements

This project was supported by National Natural Science Foundation of China (No. 62306132), Guang-

dong Basic and Applied Basic Research Foundation (No. 2025A1515011564), Natural Science Foundation of Shanghai (No. 25ZR1402136). We thank the anonymous reviewers for their insightful feedback on this work.

References

AI@Meta. 2024. Llama 3 model card.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Ale s Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Chung-Ching Chang, David Reitter, Renat Aksitov, and Yun-Hsuan Sung. 2023. Kl-divergence guided temperature sampling. *Preprint*, arXiv:2306.01286.

Yi Cheng, Xiao Liang, Yeyun Gong, Wen Xiao, Song Wang, Yuji Zhang, Wenjun Hou, Kaishuai Xu, Wenge Liu, Wenjie Li, Jian Jiao, Qi Chen, Peng CHENG, and Wayne Xiong. 2025. Integrative decoding: Improving factuality via implicit self-consistency. In *The Thirteenth International Conference on Learning Representations*.

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

DeepSeek-AI, Aixin Liu, Bei Feng, and Bing Xue. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. 2024. Length-controlled alpacaeval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*.

Esteban Garces Arias, Julian Rodemann, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher. 2024. Adaptive contrastive search: Uncertainty-guided decoding for open-ended text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15060–15080, Miami, Florida, USA. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Christoph Jansen, Georg Schollmeyer, Julian Rodemann, Hannah Blocher, and Thomas Augustin. 2024. Statistical multicriteria benchmarking via the gsd-front. *Advances in Neural Information Processing Systems*, 37:98143–98179.

Jiyeon Kim, Hyunji Lee, Hyowon Cho, Joel Jang, Hyeonbin Hwang, Seungpil Won, Youbin Ahn, Dohaeng Lee, and Minjoon Seo. 2025. Knowledge entropy decay during language model pretraining hinders new knowledge acquisition. In *The Thirteenth International Conference on Learning Representations*.

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

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

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

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

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. 2025. Preserving diversity in supervised fine-tuning of large language models. In *The Thirteenth International Conference on Learning Representations*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A. Smith. 2024a. Tuning language models by proxy. In *First Conference on Language Modeling*.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024b. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*.

Siyang Liu, Sahand Sabour, Yinhe Zheng, Pei Ke, Xiaoyan Zhu, and Minlie Huang. 2022. Rethinking and refining the distinct metric. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 762–770, Dublin, Ireland. Association for Computational Linguistics.

Avshalom Manevich and Reut Tsarfaty. 2024. Mitigating hallucinations in large vision-language models (LVLMs) via language-contrastive decoding (LCD). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6008–6022, Bangkok, Thailand. Association for Computational Linguistics.

Nguyen Nhat Minh, Andrew Baker, Clement Neo, Allen G Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. 2025. Turning up the heat: Min-p sampling for creative and coherent LLM outputs. In *The Thirteenth International Conference on Learning Representations*.

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*.

OpenAI, Josh Achiam, Steven Adler, and Sandhini Agarwal. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? *Preprint*, arXiv:2405.00492.

Jonathan Pei, Kevin Yang, and Dan Klein. 2023. PREADD: Prefix-adaptive decoding for controlled text generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10018–10037, Toronto, Canada. Association for Computational Linguistics.

Allan Raventos, Mansheej Paul, Feng Chen, and Surya Ganguli. 2023. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Matthew Renze. 2024. The effect of sampling temperature on problem solving in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7346–7356, Miami, Florida, USA. Association for Computational Linguistics.

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 2024. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 21–33, St. Julian's, Malta. Association for Computational Linguistics.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

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

Sean Welleck, Amanda Bertsch, Matthew Finlayson, Hailey Schoelkopf, Alex Xie, Graham Neubig, Ilia Kulikov, and Zaid Harchaoui. 2024. From decoding to meta-generation: Inference-time algorithms for large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jiayi Wu, Hao Sun, Hengyi Cai, Lixin Su, Shuaiqiang Wang, Dawei Yin, Xiang Li, and Ming Gao. 2024. Cross-model control: Improving multiple large language models in one-time training. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Shimao Zhang, Yu Bao, and Shujian Huang. 2024a. Edt: Improving large language models' generation by entropy-based dynamic temperature sampling. *Preprint*, arXiv:2403.14541.

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

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zhanhui Zhou, Zhixuan Liu, Jie Liu, Zhichen Dong, Chao Yang, and Yu Qiao. 2024. Weak-to-strong search: Align large language models via searching over small language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

Yuqi Zhu, Jia Li, Ge Li, YunFei Zhao, Zhi Jin, and Hong Mei. 2024. Hot or cold? adaptive temperature sampling for code generation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 437–445.

A Algorithm for G2

This appendix provides a detailed algorithmic specification of our proposed **G2** framework, which was introduced and described in Section 3. Algorithm 1 outlines the complete procedure for generating the i-th diverse response, A_i. The pseudocode illustrates the integration of G2's core components: the representative prior response sampling (detailed in Section 3.3), the selective intervention strategy (detailed in Section 3.2), and the dual Guide mechanism.

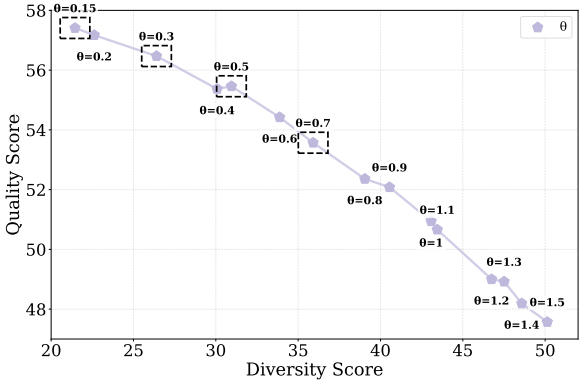


Figure 5: Impact of the intervention strength hyperparameter θ on both quality and diversity on 100 samples from the WMT'14 Fr-En validation set.

B Additional Results

B.1 Hyperparameter Searching

The hyperparameter θ in Equation 4 dictates the strength of the Guide's intervention on the base generator's token distribution. A larger value of θ signifies a stronger intervention, which typically leads to increased diversity in the generated outputs. Conversely, when $\theta = 0$, no intervention is applied by the Guides, and the generation process is equivalent to standard sampling from the base model.

To determine suitable values for θ to be employed throughout our main experiments, we conducted a systematic search on a validation set. For this purpose, we selected 100 samples from the WMT'14 French-English (Fr-En) machine translation task. The quality of the generated translations was assessed using BLEU and COMET (detailed in Section 4.1), while diversity was measured by Div-BLEU, EAD, and Sentence-Bert (detailed in Section 4.4.1).

The impact of varying θ on both diversity and quality is illustrated in Figure 5. As anticipated, an increase in θ consistently leads to higher scores across all diversity metrics. This enhancement in diversity naturally presents a trade-off with generation quality, a common consideration in developing diversifying generation methods, which our approach aims to carefully balance. Based on these observations from the validation set, we selected a specific set of θ values—0.15, 0.3, 0.5, and 0.7—for comprehensive evaluation in our main experiments. We chose not to explore θ values beyond 0.7, as preliminary results indicated that these higher intervention strengths tended to cause a more pronounced degradation in generation qual-

Algorithm 1 G2: Generating the i-th Diverse Response A_i

Require: Query Q; Base prompt $\mathcal{P}$; Set of previously generated responses $A_{<i}$; Diversity Guide prompt $\mathcal{P}^+$; Dedupe Guide prompt $\mathcal{P}^-$; Entropy threshold β; Intervention strength θ; Max generation length T_{max}; Number of representative prior responses K_{repr}.

Ensure: i-th response A_i.

$A_i \leftarrow$ empty sequence
$A'_{<i} \leftarrow \emptyset$ ▷ Initialize representative prior responses
if $i > 1$ **and** $A_{<i} \neq \emptyset$ **then**
  $A'_{<i} \leftarrow \text{RepresentativePriorResponseSampling}(A_{<i}, K_{repr})$ ▷ As per Section 3.3
for $t = 1 \rightarrow T_{\text{max}}$ **do**
  $\mathbf{x}_{<t} \leftarrow$ current tokens in A_i
  $\mathbf{z}_t \leftarrow M[\mathcal{P}, Q](\mathbf{x}_{<t})$ ▷ Base LLM logits
  if $A'_{<i} \neq \emptyset$ **then**
    $\mathbf{z}_t^+ \leftarrow M[\mathcal{P}^+, Q, A'_{<i}](\mathbf{x}_{<t})$ ▷ Diversity Guide logits
    $\mathbf{z}_t^- \leftarrow M[\mathcal{P}^-, Q, A'_{<i}](\mathbf{x}_{<t})$ ▷ Dedupe Guide logits
    $P_{\text{base},t} \leftarrow \text{softmax}(\mathbf{z}_t)$
    $H_t \leftarrow H(P_{\text{base},t})$ ▷ Entropy of base model's predictive distribution
    if $H_t \geq \beta$ **then**
      $\alpha_t \leftarrow \theta$ ▷ Activate intervention (sufficient uncertainty)
    else
      $\alpha_t \leftarrow 0$ ▷ Confident prediction, no intervention
  else
    $\alpha_t \leftarrow 0$ ▷ No prior responses to Guides (e.g., for A_1)
    $\mathbf{z}_t^+ \leftarrow \mathbf{0}; \mathbf{z}_t^- \leftarrow \mathbf{0}$ ▷ Ensure no undefined behavior if $\alpha_t = 0$
  $\mathbf{z}_t^{\text{final}} \leftarrow \mathbf{z}_t + \alpha_t(\mathbf{z}_t^+ - \mathbf{z}_t^-)$ ▷ Modulate distribution (Equation 4)
  $P_t \leftarrow \text{softmax}(\mathbf{z}_t^{\text{final}})$
  $x_t \sim P_t$ ▷ Sample next token X_t
  Append x_t to A_i
  if $x_t =$ EOS_TOKEN **then**
    break ▷ End of sequence
return A_i

ity.

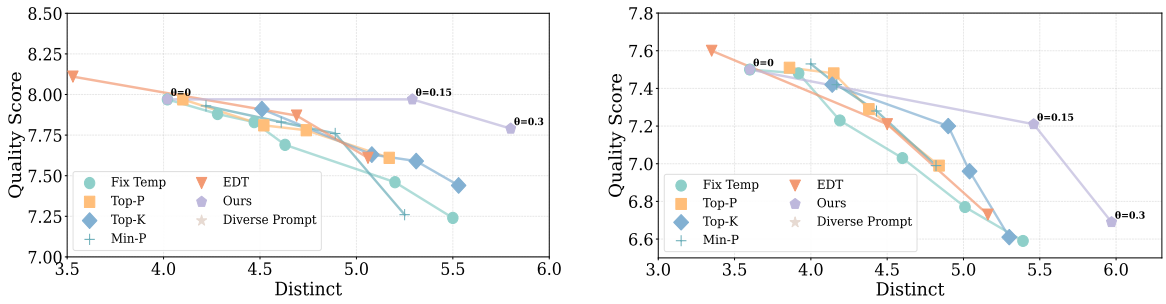
B.2 NoveltyBench

Supplementing the main text (Section 4.2.2), this section presents the complete figure results for NoveltyBench, necessitated by space limitations. This section includes Figure 6 with complete hyperparameter settings, and Table 6 providing additional hyperparameters compared to the main table. Since including all hyperparameter combinations in a table would occupy excessive space, the complete hyperparameter results are presented in the figure format.

Figure 6 visualizes the diversity-quality trade-off for G2 against baseline methods across various parameter configurations on NoveltyBench. The horizontal axis denotes diversity, while the vertical axis represents quality. For both Llama3-8B-Instruct (Figure 7a) and Qwen2.5-7B-Instruct (Figure 7b), G2 consistently populates the upper-right location. This positioning demonstrates its superior ability to enhance generation diversity while preserving quality, thereby achieving a more favorable trade-off than the compared baselines.

B.3 Instrution-Following Task

This section presents supplementary results for the instruction-following tasks in Section 4.3, including additional parameter configurations not detailed in the main text due to space limitations. Given the considerable cost associated with GPT-based evaluations, we assessed a curated subset of parameters for each baseline method. These selected parameters are specified in the legends of Figure 7. As illustrated in the figure for both AlpacaEval 2.0 and MT-Bench, G2 consistently positions itself



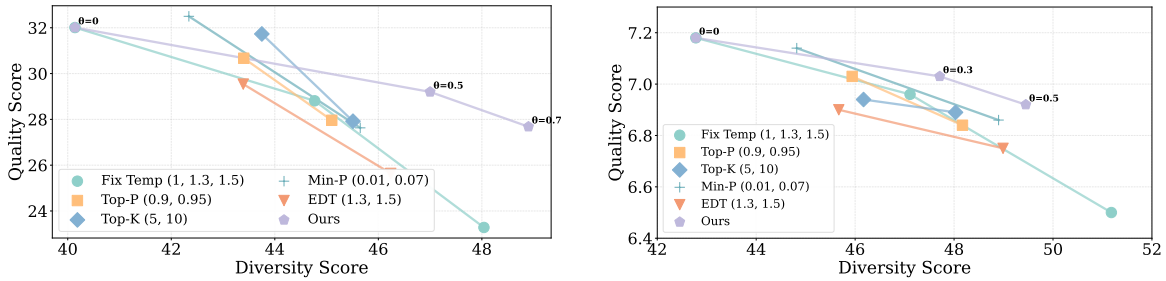
(a) Llama3-8B-Instruct

(b) Qwen2.5-7B-Instruct

Figure 6: Diversity-quality curves on NoveltyBench between G2 and other baselines under different settings.

Methods	**Div-BLEU**	**EAD**	**Sent-Bert**	**Diversity** (↑)	**Distinct** (↑)	**Quality** (↑)
Llama3-8B-Instruct	53.59	69.22	29.40	45.40	4.02	7.97
w. Temperature (T=1.3)	62.29	73.61	34.44	51.20	4.63	7.69
w. Temperature (T=1.5)	67.66	77.42	36.78	54.66	5.50	7.24
w. Top-P (T=1.5, P=0.9)	61.04	72.61	32.63	49.73	4.74	7.78
w. Top-P (T=1.5, P=0.95)	64.60	75.47	36.44	53.24	5.17	7.61
w. Top-K (T=1.5, K=5)	59.03	72.66	33.31	49.58	4.51	7.91
w. Top-K (T=1.5, K=10)	64.65	75.86	36.17	53.21	5.08	7.63
w. Min-P (T=1.5, P=0.01)	65.24	75.09	35.62	52.89	5.25	7.61
w. Min-P (T=1.5, P=0.03)	60.80	73.57	35.57	51.38	4.89	7.76
w. Min-P (T=1.5, P=0.05)	58.77	72.02	32.95	49.17	4.61	7.83
w. EDA (T=1.3, θ=0.1)	58.59	72.02	32.78	49.04	4.69	7.87
w. EDA (T=1.5, θ=0.1)	60.81	73.93	33.73	50.55	5.06	7.61
w. G2 (T=1, θ=0.15)	57.75	73.57	36.02	50.84	5.29	7.97
w. G2 (T=1, θ=0.3)	64.72	78.01	37.91	54.64	5.80	7.79
Qwen2.5-7B-Instruct	56.78	69.42	32.84	47.97	3.60	7.50
w. Temperature (T=1.3)	66.25	75.04	34.73	52.69	4.60	7.03
w. Temperature (T=1.5)	71.34	80.08	39.85	57.78	5.39	6.59
w. Top-P (T=1.5, P=0.9)	62.38	74.36	35.15	51.76	4.38	7.29
w. Top-P (T=1.5, P=0.95)	64.76	75.72	36.20	53.22	4.84	6.99
w. Top-K (T=1.5, K=5)	60.05	74.48	33.40	50.33	4.14	7.42
w. Top-K (T=1.5, K=10)	64.51	76.80	35.30	52.98	4.90	7.20
w. Min-P (T=1.5, P=0.01)	66.29	75.04	34.54	52.60	4.82	6.99
w. Min-P (T=1.5, P=0.03)	62.34	72.47	33.00	50.20	4.43	7.28
w. Min-P (T=1.5, P=0.05)	60.00	73.08	33.90	50.22	4.17	7.42
w. EDA (T=1.3, θ=0.1)	62.82	73.28	34.49	51.27	4.50	7.21
w. EDA (T=1.5, θ=0.1)	68.06	77.28	37.08	54.88	5.16	6.73
w. G2 (T=1, θ=0.15)	65.53	78.79	38.04	55.10	5.46	7.21
w. G2 (T=1, θ=0.3)	71.47	81.78	39.74	58.18	5.97	6.69

Table 6: Supplement NoveltyBench results with expanded hyperparameter configurations, supplementing Table 1 in the main text. This table provides comprehensive comparisons across all tested parameter settings for both models.



(a) AlpacaEval 2.0

(b) MT-Bench

Figure 7: Diversity-quality curves on AlpacEval 2.0 and MT-Bench between G2 and other baselines under different settings.

in the upper-right location of the diversity-quality plots. This indicates that G2 achieves a more advantageous trade-off between generation diversity and response quality compared to the evaluated baselines.

C Additional Analysis

C.1 Qualitative Analysis on Mathematical Reasoning

To provide insights into the nature of diversity obtained by G2, we conduct both quantitative and qualitative analyses on the GSM8K dataset, addressing reviewer concerns about the semantic meaningfulness of our diversity improvements.

Quantitative Analysis G2 achieves a higher Pass@3 accuracy (92.27% vs. 91.13% for the baseline) and generates more diverse answers (1.54 vs. 1.38 unique answers/question). This indicates that G2's diversity is not only greater in quantity but also more productive in helping the model arrive at correct solutions.

Qualitative Analysis Among the 239 GSM8K questions where both G2 and the baseline fail on the first attempt, we found 41 cases where the baseline method continues to fail across all three attempts, while
mname successfully recovers the correct answer by exploring alternative reasoning paths. Due to space constraints, we include one representative example below:

Question: Poppy has a 1000-piece puzzle. She places 1/4, then her mom places 1/3 of the remaining pieces. How many are left?

Baseline (Fail): In all 3 attempts, the baseline produces the same incorrect answer (503), repeatedly applying an imprecise decimal approximation (0.33) for "a third" and rounding the result incorrectly.

G2 (Success): G2 also produces 503 on its first attempt. However, on the second attempt, it is guided to a distinct and more precise reasoning path, correctly computing 750 ÷ 3 = 250, and arriving at the correct answer of 500.

This example illustrates how G2's diversity is semantically meaningful, enabling the model to recover from initial reasoning errors—an effect not captured by surface-level diversity metrics alone.

C.2 Statistical Significance Analysis

To formally assess the significance of our results, we adopted the **GSD-front framework**, following the methodology proposed in (Jansen et al., 2024). This framework provides a rigorous statistical approach for multicriteria benchmarking.

While our analysis is based on aggregated scores (yielding a deterministic empirical front), the GSD framework allows us to formally identify the set of non-dominated methods. Our results show that G2 (at $\theta = 0.15$ and $\theta = 0.3$) lies on the GSD-front, whereas many conventional baselines (Top-P, Top-K et al.) are formally dominated, indicating that their trade-offs are suboptimal across evaluation dimensions. This suggests that G2's performance improvements are statistically robust, and that the baseline decoding methods fail to achieve competitive diversity–quality trade-offs under rigorous multicriteria analysis.

C.3 Prompt Sensitivity Analysis

In this section, we conducted additional experiments to evaluate the robustness of our method across different prompt formulations.

It is important to clarify that both G2's base generator and all baseline decoding methods (e.g., temperature sampling, top-p, etc.) use the same orig-

inal task prompt, which directly corresponds to the input question or instruction. The only components that involve additional prompting are G2's two contrastive modules: the Diversity Guide and Dedupe Guide, which rely on lightweight role-based prompts.

To assess G2's robustness to prompt variation, we used a capable LLM (Gemini 2.5 Pro) to automatically generate two alternative sets of prompts for the Diversity and Dedupe guides. We then re-evaluated G2 on NoveltyBench using these new prompt formulations while keeping all other experimental conditions identical. The results demonstrate consistent performance across different prompt variations:

Prompt Variation	**Distinct** (↑)	**Quality** (↑)
Original Prompts	5.80	7.79
Prompt Set 1	5.77	7.87
Prompt Set 2	5.42	7.85

Table 7: Performance of G2 across different prompt formulations on NoveltyBench.

Despite minor variations in absolute scores, both alternative prompt sets outperform all baseline decoding methods on NoveltyBench (as shown in Table 6).

These results demonstrate that G2's effectiveness is not dependent on specific prompt phrasing, but instead stems from the contrastive guidance mechanism itself, which proves robust under moderate prompt variations. The consistent superior performance across different prompt formulations indicates that the core methodology (leveraging contrastive decoding with diversity and deduplication guides) is the primary driver of G2's improvements, rather than careful prompt engineering.

D Prompts

In this section, we will provide the prompts used by the basic generator, positive Guide, and negative Guide across various datasets. For NoveltyBench, given its strong emphasis on diversity, the experiment employs more forceful and detailed guiding prompts. These are exemplified by the prompt designed to elicit new solutions (as shown in Figure 8) and the prompt that demands exact replication (as shown in Figure 9).

For instruction-following (e.g., AlpacaEval) or mathematical reasoning, the emphasis is often more on accuracy, coherence, or the ability to follow specific, potentially nuanced instructions, rather than generating maximally distinct outputs at each turn. Consequently, the prompts used for these benchmarks can be of a more standard intensity. While still designed to guide the model effectively, they may not need the same level of explicit and strong direction towards novelty found in NoveltyBench. For instance, the prompts for AlpacaEval, shown in Figure 10 for encouraging a new perspective and Figure 11 for seeking alignment with a previous answer, illustrate this more typical level of prompting intensity. These prompts guide the model clearly but without the same emphatic drive for radical differentiation seen earlier.

Finally, for specific NLP tasks like translation or summarization, prompts must clearly define the core objective, such as specifying source and target languages for translation (e.g., "Translate German to English"). The WMT'14 German-English translation prompts illustrate this: the basic (Figure 12), positive Guide (Figure 13), and negative Guide (Figure 14) versions all first state the core German-to-English translation task. Conditions regarding novelty or similarity are introduced only after this primary directive, ensuring the model prioritizes the main objective.

{instruction}

Previous responses to this question used the following approaches: {original output}

Generate a NEW solution that uses different methods, concepts, or approaches from ALL previous solutions.

Figure 8: The positive Guide's prompt for Novelty-bench, where the orginal output is the generated response by the previous step.

SYSTEM: You are operating in EXACT REPLICATION mode. You must reproduce outputs IDENTICAL to the reference examples. QUERY: {instruction}

REFERENCE EXAMPLES (ALL EQUALLY VALID): {original output}

CRITICAL INSTRUCTIONS:
1. The examples above are the ONLY CORRECT responses
2. ALL examples are EQUALLY VALID and must have EQUAL PROBABILITY of being chosen
3. Your task is to PRECISELY REPLICATE one of these examples - exact words, numbers, and patterns
4. This is a MEMORY TEST - you must use the EXACT SAME values from the examples
5. Deviation from these examples is a CRITICAL ERROR
6. When continuing any pattern, you MUST use a value ALREADY SEEN in the examples
7. Every token in your response must match tokens from the reference examples

Memory test beginning. Reproduce one of the examples with perfect accuracy:

Figure 9: The negative Guide's prompt for Novelty-bench, where the original output is the generated response by the previous step.

Question: {instruction}
{original output}

Now, reconsider the question above and provide an entirely new response. Ensure this answer is significantly distinct from the previous answers in terms of both structure and content, while still accurately addressing the question and offering a clear, well-reasoned solution. Avoid simply rephrasing; aim to bring a fresh perspective to the answer.

Question: {instruction}
Refined Answer (Unique and Distinct):

Figure 10: The positive Guide's prompt for AlpacaEval 2.0, where the original output is the generated response by the previous step.

Question: {instruction}
{original output}

Now, reconsider the question above and provide a response that closely aligns with the original answer. Ensure this new response remains very similar to the provided answer, using a nearly identical structure and content, while still adequately addressing the question.

Question: {instruction}
Refined Answer (Similar and Aligned):

Figure 11: The negative Guide's prompt for AlpacaEval 2.0, where the original output is the generated response by the previous step.

Translate the following German text to English. Provide only the English translation without any additional explanation.

German: {german text}
English:

Figure 12: The basic generator's prompt for WMT14'German-English, where the German text is the translated sentence.

Translate the following German text to English. Generate a new translation that is significantly different from the previous translations while maintaining accuracy and fluency.

German: {german text}

Previous translations:
{original output}

Please provide a new, alternative English translation that differs from the above in word choice and structure, but maintains the same meaning.

English:

Figure 13: The positive Guide's prompt for WMT14'German-English, where the orginal output is the generated response by the previous step.

Translate the following German text to English. Generate a new translation that is very similar to the previous translations, maintaining the same word choices and sentence structure whenever possible.

German: {german text}

Previous translations:
{original output}

Instructions:
- Study the patterns and word choices in the previous translations carefully
- Use the same vocabulary and phrasing as much as possible - Keep the sentence structure highly similar
- Only make minimal necessary adjustments for fluency

Please provide a new English translation that closely aligns with the previous versions:

Figure 14: The negative Guide's prompt for WMT14'German-English, where the original output is the generated response by the previous step.