
DIVERSED: Relaxed Speculative Decoding via Dynamic Ensemble Verification

Ziyi Wang¹, Siva Rajesh Kasa², Ankith M S², Santhosh Kumar Kasa², Jiaru Zou³,
Nan Jiang⁴, Sumit Negi², Ruqi Zhang¹, Qifan Song¹

¹Purdue University, ²Amazon.com Inc., ³University of Illinois Urbana-Champaign,

⁴University of Texas at El Paso

{wang4538, ruqiz, qfsong}@purdue.edu, jiaruz2@illinois.edu
{kasasiva, ankiths, suminegi}@amazon.com, njiang@utep.edu

Abstract

Speculative decoding is an effective technique for accelerating large language model (LLM) inference by drafting multiple tokens in parallel. However, its practical speedup is often limited by a rigid verification step, which strictly enforces that the accepted token distribution exactly matches that of the target model. This constraint leads to the rejection of many plausible tokens, reducing the acceptance rate and limiting overall time speedup. To overcome this limitation, we propose a relaxed verification framework: DynamIc VERification RELaxed SpECulative Decoding (DIVERSED), that improves efficiency while preserving generation quality. DIVERSED introduces a learned ensemble-based verifier that blends the draft and target model distributions using dynamic mixing weights. This mixture distribution serves as a more flexible verification target, increasing token acceptance without compromising on overall correctness. We provide theoretical justification for our approach and demonstrate empirically that DIVERSED achieves significantly higher inference efficiency compared to traditional speculative decoding methods¹.

1 Introduction

Recent advances in large language models (LLMs) increasingly leverage additional inference-time computation to address demanding tasks, ranging from multi-step mathematics [Kojima et al., 2022, Wei et al., 2022] to complex software development [Rozière et al., 2023, Yan et al., 2025, Jiang et al., 2024] and tool use [Schick et al., 2023, Qin et al., 2024]. These techniques have steadily improved model performance. However, the computational cost of such models remains substantial, making inference efficiency a critical bottleneck. Existing work has focused on optimizing attention kernels [Dao et al., 2022], model compression [Dettmers et al., 2022], and accelerating decoding [Leviathan et al., 2023, Narasimhan et al., 2025].

Among decoding strategies, speculative decoding has recently emerged as a particularly effective approach [Leviathan et al., 2023, Chen et al., 2023]. It employs a small draft model to propose tokens and then uses a large target model to verify the tokens in parallel, achieving target-level quality with reduced latency. Subsequent research has developed along two main directions: (1) lossless architectures, which preserve exact agreement with the target, and (2) lossy variants, which accept more draft tokens and may compromise generation quality. The first line of work includes self-speculative architectures that restructure or internalize the draft process [Cai et al., 2024, Li et al., 2024b, Elhoushi et al., 2024], and alignment strategies that train draft models to produce more target-like outputs, thereby increasing acceptance [Zhou et al., 2024, Goel et al., 2024, Liu et al., 2023]. On the other hand, lossy approaches intentionally relax token-level agreement with the

¹Our code is built on Hugging Face Transformers and integrates easily into existing generation pipelines.

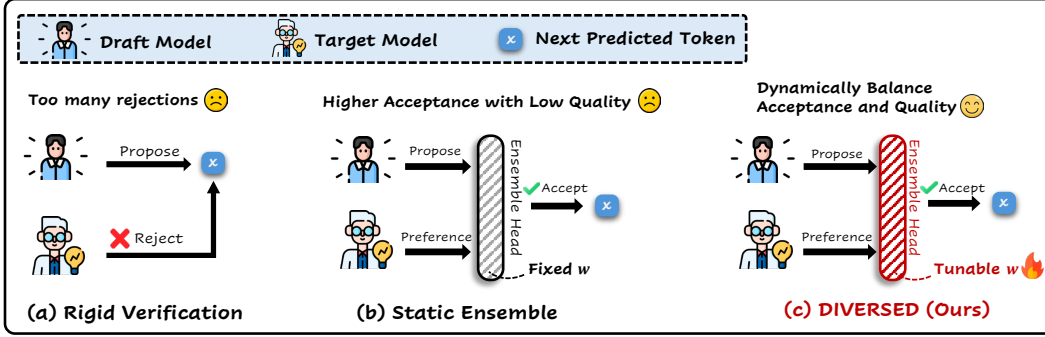


Figure 1: Comparison of (a) classic speculative decoding (“rigid verification”), (b) a static ensemble of the draft and target models, and (c) our proposed DIVERSED, which adaptively chooses ensemble coefficients at each decoding step. Compared with rigid verification, our DIVERSED achieves a higher acceptance rate and similar accuracy while being faster; Compared with the static ensemble, DIVERSED yields higher accuracy.

target, trading bounded bias (causing performance degradation) for higher acceptance and throughput [Bachmann et al., 2025, Yin et al., 2024]. Yin et al. [2024] theoretically characterizes the tradeoffs between acceptance and quality: the more acceptance draft tokens, the worse the generation quality.

An unsolved challenge of speculative decoding is that rigid verification discards many high-quality draft tokens simply because they are not the target model’s exact preference, leading to wasted computation [Yan et al., 2025, Bachmann et al., 2025]. An intuitive solution is *static ensemble*, a simple, training-free verification rule that *explicitly solves* the Pareto optimal outlined by Yin et al. [2024]. The static ensemble could interpolate along the acceptance-quality trade-off by varying the ensemble weight. However, static ensembles still exhibit low generation quality and require manually chosen coefficients that vary across datasets and tasks (e.g., code generation vs. math). In

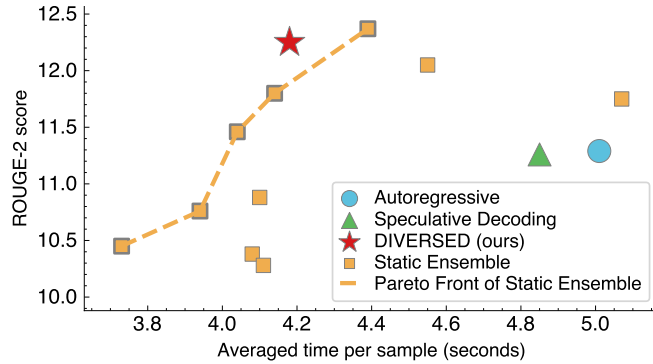


Figure 2: Our method outperforms the Pareto front of static ensemble and other baselines. Results are reported on the CNNDM dataset using the target/draft model: Llama-3.1-8B-Instruct and Llama-3.2-1B-Instruct, with temperature be 0.

Figure 3, we show a case where the Static ensemble could hurt the generation quality.

To address this limitation, we propose **D**ynam**I**c **V**erification **R**elaxed **S**peculative **D**ecoding (DIVERSED), a dynamic ensemble verifier that mixes the target and draft distributions with mixing token-dependent weights. The weights are learned to maximize quality under a high-acceptance constraint. In Figure 2, DIVERSED attains higher acceptance at comparable quality outperforming the Pareto front obtained by the static ensemble.

Theoretically, we prove that the *static ensemble* verifier lies on the acceptance-quality Pareto front proposed by Yin et al. [2024]. Importantly, this reveals that practitioners do not need to maintain a fleet of draft–target pairs for different user requirements: a *single pair of models* can flexibly serve diverse latency–quality trade-offs simply by adjusting the static ensemble weight. We also derive an exact, step-dependent expression for the expected accepted length that captures position effects ignored by prior analyses, offering a practical metric to predict speedup and guide verifier design.

Experimentally, we show DIVERSED delivers substantially higher acceptance rates, while preserving task quality across all model pairs and benchmarks. We further show that wall-clock speedup closely correlated to acceptance rate T_N/N : increases in acceptance rate reliably translate into end-to-end latency reductions. Finally, we observe a clear efficiency–quality Pareto front, and DIVERSED consistently attains a strong balance between the two.

2 Preliminary

Speculative decoding. In autoregressive LLMs, a target model with distribution $p(\cdot \mid x_{0:t-1})$ generates tokens sequentially from a prompt x_0 , producing $x_1, \dots, x_T$ one at a time. Speculative decoding augments this process with a smaller, faster draft model $q(\cdot \mid x_{0:t-1})$ that proposes a block of N draft tokens $\tilde{x}_{t:t+N-1}$. For notational convenience, we write $p_t(\cdot) = p(\cdot \mid x_{0:t-1})$ and $q_t(\cdot) = q(\cdot \mid x_{0:t-1})$. The target model then verifies in parallel by computing $p(\tilde{x}_{t:t+N-1} \mid x_{0:t-1})$ and accepts $\tilde{x}_t$ with probability

$$b_t(\tilde{x}_t) = \min \left\{ 1, \frac{p_t(\tilde{x}_t)}{q_t(\tilde{x}_t)} \right\}.$$

If accepted, we set x_t as $\tilde{x}_t$. Otherwise, the target samples a correction token x_t from the fallback distribution: $\frac{[p_t(x) - q_t(x)]_+}{\sum_z [p_t(z) - q_t(z)]_+}$, where $[x]_+ = \max\{0, x\}$. Under standard conditions, the resulting stream of tokens is exactly distributed according to the target model [Leviathan et al., 2023].

Let T_N be the number of accepted tokens with the draft length N . Prior work [Leviathan et al., 2023] has shown that $\mathbb{E}[T_N] = \frac{1 - \beta^{N+1}}{1 - \beta}$, where β is the expected accepted probability $\mathbb{E}_{\tilde{x} \sim q}[\alpha(\tilde{x})] = 1 - \text{TV}(p, q)$. However, these results are derived under the independent and identical (i.i.d.) token verification assumption and have not considered the effect of the draft token positions in the draft sequence.

Lossy speculative decoding and its Pareto Front. A lossy variant accepts a drafted token $\tilde{x}_t$ with a more general accept rule $b_t(\tilde{x}_t)$ and, if rejected, samples from a fallback $P_t(\cdot)$, inducing a token distribution $\mathbb{P}_{b_t}^{P_t}$; the detailed algorithm can be found in the Algorithm 1 in Yin et al. [2024]. The accept rule must have the nature such that $b_t(\tilde{x}_t) > \alpha(\tilde{x}_t)$ to accept more tokens. The design objective of lossy speculative decoding is a multi-objective problem that simultaneously minimizes the rejection probability $\mathbb{P}_{b_t}(\text{Reject})$ and the total variation bias $\text{TV}(\mathbb{P}_{b_t}^{P_t}, p)$. We define Pareto optimality to characterize this trade-off between acceptance and distribution bias in TV.

Definition 2.1 (Pareto front solution for lossy speculative decoding). *An accept rule b_t with fallback P_t dominates b'_t with P'_t if $\mathbb{P}_{b_t}(\text{Reject}) < \mathbb{P}_{b'_t}(\text{Reject})$ and $\text{TV}(\mathbb{P}_{b_t}^{P_t}, p) < \text{TV}(\mathbb{P}_{b'_t}^{P'_t}, p)$. A pair (b_t, P_t) is Pareto-optimal (non-dominated) if no other pair dominates it; the set of all such pairs is the Pareto front.*

Theorem 2.2 (Yin et al. [2024], Theorem 5; informal). *Any (b_t, P_t) achieves Pareto-optimal if and only if (b_t, P_t) satisfies*

$$\mathbb{P}_{b_t}(\text{Reject}) + \text{TV}(\mathbb{P}_{b_t}^{P_t}, p) = \text{TV}(q, p). \quad (1)$$

Theorem 2.2 implies that reducing the total number of rejections—thereby improving time efficiency—inevitably introduces bias.

3 Methodology

3.1 Static Ensemble

Motivating example. Figure 4 (a) examines the relation between wall-clock time and acceptance rate, using data from generating 128 tokens on the CNNDM dataset [See et al., 2017], across target/draft pairs from the Llama, Qwen, and Gemma families. For each pair, we observe that wall-clock time decreases monotonically as acceptance rate increases. This indicates that raising the acceptance rate is the key to wall-clock time speedup.

Thus, a simple yet effective method to improve the acceptance rate is *static ensemble*, which leverages a fixed ensemble of the target distribution $p_t(\cdot)$ and the draft distribution $q_t(\cdot)$ as the verification distribution. Given a constant mixing weight $0 < w \leq 1$, we define the ensemble distribution ν_t as:

$$\nu_t(\cdot) = wp_t(\cdot) + (1 - w)q_t(\cdot). \quad (2)$$

[Prompt] Each bird eats 12 beetles per day, ..., how many beetles are eaten each day? Each jaguar eats 5 snakes per day, so 6 jaguars will eat 6 ... per day. Each bird eats 12 beetles [and] there ...✓ [So], each bird...✓ day, and there are 90 birds, so in total they eat $12 * 90 = 1080$ beetles per day. [90]birds eat $12 * 90 = 1080$ beetles per day. ✓
[Prompt] Alexis is applying for a new job ... How much did Alexis pay for the shoes? Alexis spent $\\$30 + \dots + \\$18 = \\$143$ on the items she has receipts for. She has \$16 left from her [\$153] on the ... she spent \$49. ✗ budget, so she spent ... \$184 in total. Therefore, she spent $\\$184 - \\$143 = \\$41$ on the shoes. [on] the shoes....✗

Figure 3: Draft–target token mismatch analysis in speculative decoding verification. Black marks tokens both models agree on; **green** marks accepted mismatches that still yield the correct answer; **red** marks accepted mismatches that lead to an incorrect answer.

During the verification stage, a draft token $\tilde{x}_t$ is accepted with probability $b_{\nu_t}(\tilde{x}_t) = \min\{1, \frac{\nu_t(\tilde{x}_t)}{q_t(\tilde{x}_t)}\}$. If the token is rejected, we resample x_t from the fallback distribution:

$$P_{\nu_t}(x) = \frac{[\nu_t(x) - q_t(x)]_+}{\sum_z [\nu_t(z) - q_t(z)]_+}.$$

Following prior analysis [Chen et al., 2023, Leviathan et al., 2023], we can show that this verification and resampling mechanism ensures that the resulting token distribution x_t matches the intended ensemble distribution ν_t .

Proposition 3.1. *The static ensemble method achieves the Pareto front, i.e., (ν_t, P_{ν_t}) satisfies:*

$$\mathbb{P}_{\nu_t}(\text{reject}) + \text{TV}(\nu_t, p) = \text{TV}(q, p).$$

Proof is given in Appendix A.4.

Together with Theorem 2.2, proposition 3.1 demonstrates that static ensemble relaxation exactly characterizes the Pareto-optimal between acceptance probability and distributional bias. By varying the ensemble weight w , the method traverses the entire Pareto front.

In the next lemma, we show that our static ensemble verification could improve the probability of accepting a single draft token $\tilde{x}_t$.

Lemma 3.2. *Let draft and target models with distributions q_t and p_t . For a constant $0 < w \leq 1$, define the accept rule $b_{\nu_t}(x) = \min\{1, \frac{\nu_t(x)}{q_t(x)}\}$. Then*

$$\mathbb{P}_{\nu_t}(\tilde{x}_t \text{ accepted} \mid x_{0:t-1}) = 1 - w \text{TV}(q, p).$$

The proof is given in Appendix A.2.

Lemma 3.2 states that the Static Ensemble increases the acceptance probability from $1 - \text{TV}(q, p)$ to $1 - w \text{TV}(q, p)$. When $w = 1$, we recover the lossless speculative decoding; when $w < 1$, relaxation boosts acceptance.

3.2 DIVERSED: Dynamic VERification RELaxed SpECulative Decoding

Motivating example. Take Figure 3 as an example. We analyze how relaxation on different tokens affects the generation quality. Some tokens tolerate relaxation without degrading quality, whereas relaxing others harms quality, indicating that relaxation should be context dependent. For tokens that strongly influence quality (e.g., a token that sends a math solution off course), the weights should lean toward the target distribution; for less influential tokens, the verifier can prefer accepting the draft token even under disagreement to increase acceptance.

We design a dynamic weighting scheme where, at each time step t , the weight w_t is conditioned on the context $x_{0:t-1}$. To capture this context, we build an ensemble head on top of draft and target models, leveraging both the draft model hidden representation h_t^q and the target model hidden representation

Algorithm 1 General Framework for Lossy Speculative Decoding.

Inputs: Horizon T ; draft length N ; prompt x_0 ; draft model distributions $\{q_t\}$; target model distributions $\{p_t\}$; acceptance rule $b_t : \mathcal{V} \rightarrow [0, 1]$; fallback distribution P_t .

```
1:  $n \leftarrow 1$ ;  
2: while  $n \leq T$  do  
3:   for  $t = n, \dots, n + N$  do  
4:     sample  $\tilde{x}_t \sim q_t(\cdot \mid x_{0:n-1}, \tilde{x}_{n:t-1})$  and save  $q_t$  into cache. ▷ draft rollout  
5:   for  $t = n, \dots, T$  do  
6:     compute  $p_t(\cdot \mid x_{0:n-1}, \tilde{x}_{n:t-1})$ . ▷ target parallel verification (in section 3)  
7:   for  $t = n$  to  $n + N$  do  
8:     Draw  $u \sim \text{Uniform}(0, 1)$ .  
9:     if  $u \leq b_t(\tilde{x}_t \mid x_{1:n-1}, \tilde{x}_{n:t-1})$  then  
10:      accept  $\tilde{x}_t$ ;  $x_n \leftarrow \tilde{x}_t$ ;  $n \leftarrow n + 1$ .  
11:    else  
12:      reject; sample  $x_n \sim P_t$ ;  $n \leftarrow n + 1$ ; break. ▷ discard  $\tilde{x}_{t+1:T}$  and resume  
13: return Sequence  $x_{1:T}$  sampled from  $\mathbb{P}_{b_t}^{P_t}$  distribution.
```

h_t^p , which together summarize $x_{0:t-1}$. The ensemble weight is parameterized by a simple fully connected network:

$$w_t = f_\theta(h_t^q, h_t^p).$$

The resulting dynamic ensemble distribution is then defined as

$$\nu_t^\theta(x) = w_t p_t(x) + (1 - w_t) q_t(x), \quad (3)$$

which is used during verification: a draft token is accepted with probability $\min\{1, \frac{\nu_t^\theta(x)}{q_t(x)}\}$. Theoretically, this distribution is consistent with $\mathbb{P}(x_t = x \mid x_{0:t-1}) = \nu_t^\theta(x)$, as discussed in Section 3.1.

Training. To ensure high generation quality, we train the ensemble head by treating it as a policy model. During training, the draft and target models process the same prompt, and the ensemble distribution ν_t^θ is computed as equation (3), from which the next token x_t is sampled. A challenge arises because token-level supervision is unavailable [Bachmann et al., 2025]. We therefore frame the training of ν^θ as a reinforcement learning problem. The model generates full sequences, which are evaluated with task-dependent outcome rewards R evaluated on the whole generation $x_{1:T}$. For example, in mathematical reasoning tasks, the reward R is 1 if the final answer is correct and 0 otherwise. Thus, for some distribution P over the prompt x_0 , we can formulate the problem as a constrained optimization:

$$\max_{\theta} \mathbb{E}_{x_0 \sim P, x_{1:T} \sim \nu^\theta(x_0)} [R(x_{1:T})], \quad \text{s.t.} \sum_t^T (1 - \text{TV}(q_t, \nu_t^\theta)) \geq B,$$

for some positive margin $B > 0$.

Equivalently, one can minimize an unconstrained optimization problem with a suitable $\lambda > 0$, where $J(\theta)$ is defined as

$$\mathbb{E}_{x_0 \sim P, x_{1:T} \sim \nu^\theta} \left[R(x_{1:T}) + \lambda \sum_{t=0}^T (1 - \text{TV}(q_t, \nu_t^\theta)) \right]. \quad (4)$$

To optimize the objective (4), we follow the recent REINFORCE++ algorithm [Hu, 2025], which stabilizes training through clipping and global batch averaging.

Inference. During the inference stage, the draft model autoregressively generates N draft tokens $\tilde{x}_{t:t+N-1}$ along with hidden states $h_{t:t+N-1}^q$ and distributions $q_{t:t+N-1}$. The target model processes these tokens in parallel to produce $h_{t+1:t+N}^p$ and $p_{t:t+N-1}$. The ensemble head then computes weights and distributions in parallel:

$$\begin{aligned} w_{t:t+N-1} &= f_\theta(h_{t:t+N-1}^q, h_{t:t+N-1}^p), \\ \nu_{t:t+N-1}^\theta(\cdot) &= w_{t:t+N-1} p_{t:t+N-1}(\cdot) + (1 - w_{t:t+N-1}) q_{t:t+N-1}(\cdot). \end{aligned}$$

Thus, the relaxed accept rule and fallback distribution are computed as follows:

$$b_{\nu_t^\theta}(\tilde{x}_t) = \min \left\{ 1, \frac{\nu_t^\theta(\tilde{x}_t)}{q_t(\tilde{x}_t)} \right\}, \quad P_{\nu_t^\theta}(x) = \frac{[\nu_t^\theta(x) - q_t(x)]_+}{\sum_z [\nu_t^\theta(z) - q_t(z)]_+}. \quad (5)$$

The main training and inference pipeline is summarized in Figure 6 and Algorithm 2 in Appendix B.

4 Theoretical Analysis

In this section, we derive the exact expected acceptance rate given a draft length N . Prior work [Leviathan et al., 2023] shows that the key factor driving speedup in speculative decoding is the number of accepted draft tokens, as each accepted token avoids an expensive call to the target model. In Leviathan et al. [2023], this was only approximated under an *i.i.d.* assumption on acceptance probabilities. However, this is not true in the real world; the probability of accepting a draft token is highly related to the position of the token in the draft sequence [Huang et al., 2024, Mamou et al., 2024]. Here, we provide a precise characterization of the expected acceptance length.

Assume we are in the first round, where the draft model needs to draw a sequence $\hat{x}_{1:N}$ from its probability distribution q given the prompt x_0 . Then the verifier evaluates whether to accept or reject the t -step token. Let A_t be the acceptance indicator, where $A_t = 1$ if the draft token $\tilde{x}_t$ is accepted given history $x_{0:t-1}$, and $A_t = 0$ otherwise. We define the acceptance length T_N as the position of the first rejection:

$$T_N = \min\{t \in [1, \dots, N] : A_t = 0\}.$$

When all tokens are accepted, we would have $T_N = N + 1$. Hence, $T_N - 1$ represents the number of consecutive draft tokens accepted before the first rejection. We first give the distribution of the accepted tokens, which will be used for computing the expected accepted length.

Lemma 4.1 (Distribution of Accepted Tokens). *Consider a draft model M_q and a target model M_p with conditional distributions $q(\cdot | x_{0:t-1})$ and $p(\cdot | x_{0:t-1})$ with prompt x_0 . Given that a token is accepted by speculative decoding, its distribution follows the min-kernel:*

$$\mathbb{P}(x_t = x \mid x_t \text{ is accepted}) = \frac{\min\{q(x \mid x_{0:t-1}), p(x \mid x_{0:t-1})\}}{1 - \text{TV}(q(\cdot \mid x_{0:t-1}), p(\cdot \mid x_{0:t-1}))} := \tilde{m}(x \mid x_{0:t-1})$$

Lemma 4.1 shows that accepted tokens are not distributed according to the target distribution, but instead follow the min-kernel $\tilde{m}_t(x)$. For example, if the draft generates $\tilde{x}_{1:N}$ and $x_{1:t} = \tilde{x}_{1:t}$ are all accepted, then $x_{1:t}$ comes from $\tilde{m}$. This is important when characterizing the acceptance rate, as the precondition of draft token $\tilde{x}_t$ being accepted is only when all previous drafts $\tilde{x}_{1:t-1}$ are accepted. We conduct more experiments to illustrate Lemma 4.1; the details can be found in the Appendix C.3.

Theorem 4.2 (Expected Acceptance Length). *Consider draft model M_q and target model M_p with conditional distributions $q(\cdot | x_0)$, $p(\cdot | x_0)$ and draft length N . Then the expected acceptance length $\mathbb{E}[T_N]$ is*

$$\sum_{t=1}^N \mathbb{E}_{x_{1:t-1} \sim \tilde{m}} \left[\prod_{i=1}^{t+1} (1 - \text{TV}(q(\cdot | x_{0:i-1}), p(\cdot | x_{0:i-1}))) \right].$$

Proof Sketch. By tower law we directly characterize $\mathbb{E}[T_N \mid \tilde{x}_{1:N}]$. As the fact that T_N is non-negative, we know that $\mathbb{E}[T_N \mid \tilde{x}_{1:N}] = \sum_{t=1}^N \mathbb{P}(T_N > t \mid \tilde{x}_{1:t-1})$, and $\mathbb{P}(T_N > t \mid \tilde{x}_{1:t-1})$ is the probability that the first rejection happens after the step t given a draft sequence $\tilde{x}_{1:t-1}$. We then apply the induction rule to compute $\mathbb{P}(T_N > t \mid \tilde{x}_{1:t-1})$, as it can be written as $P(\text{Accept } \tilde{x}_t \mid \tilde{x}_{1:t-1} \text{ Accepted}) \times P(\text{Accept } \tilde{x}_{1:t-1} \mid \tilde{x}_{1:t-1})$, where the min-kernel distribution $\tilde{m}$ is used to characterize the accepted tokens. $\square$

Remark 4.3. *Previous work [Yin et al., 2024] has proved that the number of rejections (i.e., tokens to be generated by the target model) equals to $\sum_{t=1}^N \mathbb{E}_{x_{1:t-1} \sim q}[\text{TV}(q(\cdot | x_{0:t-1}), p(\cdot | x_{0:t-1}))]$, and claimed that the computing cost is proportional to it. However, this view ignores the position of rejections by assuming the draft model cost is negligible. In practice, a rejection at the first draft token and a rejection at the $(N - 1)$ -th token have very different costs: in the former case, generating a block of N proposals requires at least $2N - 1$ draft calls plus one target call, whereas in the latter*

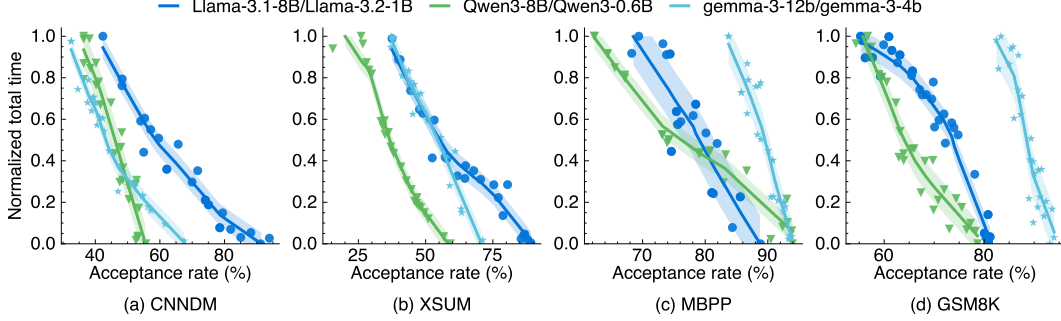


Figure 4: Results on the CNNDM, XSUM, MBPP, and GSM8k datasets for three target/draft pairs: normalized total time (lower is better) versus acceptance rate (higher is better) under speculative decoding with $N=5$. Model pairs with higher acceptance achieve lower decoding time.

it needs only N draft calls and one target call. Yin et al. [2024] treats both cases as equal cost because it counts only target invocations, yet the acceptance lengths differ markedly ($T_N = 0$ versus $T_N = N - 2$), leading to different wall-clock times.

Theorem 4.2 characterizes the expected accepted draft length using sequential dependent draft and distribution $\{q_t\}$ and $\{p_t\}$ for speculative decoding on sequence length N . Moreover, as $1 - \text{TV}(q_t, p_t) \leq 1$, the probability of accepting the draft token $\hat{x}_t$ exponentially decrease with the time step t .

5 Related Works

Speculative decoding accelerates autoregressive generation by having a small *draft* model propose tokens that a larger *target* model verifies in parallel. Leviathan et al. [2023] introduced lossless SD with exact target-distribution matching. These methods achieve about $2\times$ to $3\times$ wall-clock gains but expose a verification bottleneck. EAGLE [Li et al., 2024b], Medusa [Cai et al., 2024], and LayerSkip [Elhoushi et al., 2024] restructure the target model to reuse its computation for drafting, reducing the overhead of a separate draft model. DISCO [Mamou et al., 2024] and SpecDec++ [Huang et al., 2024] adapt the draft length dynamically, avoiding low acceptance when sequences are too long. Speculative Cascades [Narasimhan et al., 2025] unify cascaded inference with SD to trade off quality and cost. Analyses include a broad empirical study of SD bottlenecks [Yan et al., 2025] and a theoretical treatment that formalizes the efficiency–quality trade-off [Yin et al., 2024].

Improving acceptance and adaptive drafting. EAGLE [Li et al., 2024b] rethinks feature uncertainty for speculative sampling, while Medusa [Cai et al., 2024] equips LLMs with multiple decoding heads to explore continuation trees jointly. Dynamic lookahead methods such as DISCO [Mamou et al., 2024] and SpecDec++ [Huang et al., 2024] adaptively choose draft lengths, proving threshold-type optimal policies and yielding 7–11% further speedups. Yan et al. [2025] provides a systematic empirical study of bottlenecks across 350 experiments. SpecInfer [Miao et al., 2024] introduces token-tree verification for parallel throughput. SSSD and EMS-SD [Ni et al., 2025] extend SD to continuous batching and multi-sample inference, respectively, avoiding padding overhead. Speculative cascades [Narasimhan et al., 2025] unify cascaded inference with SD to trade off quality and cost. Draft & Verify [Zhang et al., 2024] eliminates the separate draft model by skipping intermediate layers. LayerSkip [Elhoushi et al., 2024] similarly reuses early layers for drafting and later layers for verification, reducing memory while maintaining fidelity. NEST [Li et al., 2024a] augments SD with nearest-neighbor retrieval to improve factuality and attribution, while retaining acceleration benefits. Yin et al. [2024] formalize SD under a Markov abstraction, deriving a linear Pareto frontier between efficiency and fidelity. This theory motivates the exploration of relaxed verification. Recent work questions the need for exact alignment. Judge Decoding [Bachmann et al., 2025] employs an LLM-as-judge to accept semantically valid but non-identical tokens. Our work, DIVERSED, advances this line by proposing a *dynamic ensemble verifier* that adaptively blends target and draft distributions, going beyond the Pareto front obtained by a static ensemble.

Table 1: Reported acceptance rate and generation quality. Columns are grouped by task. Each block compares three baselines and our method for a given *Target/Draft* pair. All results use temperature 1 and draft length $N = 5$. Our DIVERSED enjoys a higher acceptance rate and also maintains comparable quality to baselines.

Method	GSM8K		CNNDM		XSum		MBPP	
	Accept Rate ($\uparrow$)	Quality Acc ($\uparrow$)	Accept Rate($\uparrow$)	Quality ROUGE-2 ($\uparrow$)	Accept Rate($\uparrow$)	Quality ROUGE-2 ($\uparrow$)	Accept Rate($\uparrow$)	Quality pass@1 ($\uparrow$)
Autoregressive	NA	67%	NA	9.86	NA	7.03	NA	53%
SD	44.60%	67%	21.60%	9.46	20.44%	7.09	26.30%	53%
SD (Lossy)	59.81%	66%	38.86%	10.51	40.91%	7.96	66.75%	49%
SpecCascade	61.53%	67%	47.29%	11.74	43.95%	7.61	73.92%	52%
Static Ensemble	69.49%	66%	61.06%	11.46	51.58%	7.22	68.70%	52%
DIVERSED (ours)	72.61%	67%	69.96%	12.11	70.53%	7.23	85.03%	53%
(a) Target/Draft model pair is Llama-3.1-8B/Llama-3.2-1B.								
Autoregressive	NA	90%	NA	9.97	NA	4.90	NA	55%
SD	59.58%	91%	35.26%	9.85	15.60%	4.95	58.72%	55%
SD (lossy)	60.10%	86%	36.40%	10.53	19.56%	4.90	60.85%	53%
SpecCascade	58.65%	86%	38.13%	10.85	19.45%	4.97	71.27%	56%
Static Ensemble	67.58%	86%	41.07%	10.88	21.24%	4.98	76.11%	55%
DIVERSED (ours)	76.48%	88%	46.59%	10.97	47.01%	5.01	81.46%	59%
(b) Target/Draft model pair is Qwen3-8B/Qwen3-0.6B.								
Autoregressive	NA	93%	NA	9.01	NA	8.31	NA	68%
SD	84.15%	92%	40.39%	9.06	35.76%	8.27	83.25%	67%
SD (Lossy)	85.02%	90%	45.43%	10.62	39.48%	8.38	86.69%	65%
SpecCascade	84.43%	92%	51.44%	10.42	39.42%	8.27	83.76%	67%
Static Ensemble	87.62%	91%	54.48%	10.84	61.42%	7.90	86.46%	66%
DIVERSED (ours)	90.70%	92%	66.90%	10.86	63.38%	7.22	90.23%	67%
(c) Target/Draft model pair is Gemma-3-12B/Gemma-3-4B.								

6 Experiments

In this section, we show DIVERSED delivers substantially higher acceptance rates, while preserving task quality, across all model pairs and benchmarks. We further show that wall-clock speedup closely correlated to acceptance rate T_N/N : increases in acceptance rate reliably translate into end-to-end latency reductions. Finally, we observe a clear efficiency–quality Pareto front, and DIVERSED consistently attains a strong balance between the two.

6.1 Experimental Settings

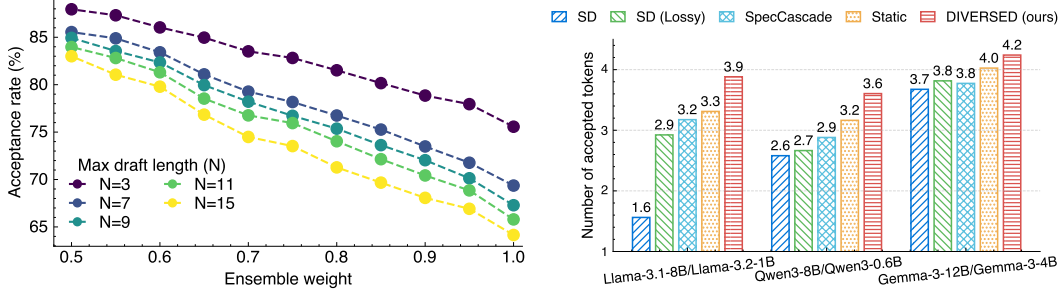
LLM setting and Datasets. We evaluate three target/draft model pairs: (i) Llama-3.1-8B-Instruct/Llama-3.2-1B-Instruct [Dubey et al., 2024], (ii) Qwen3-8B/Qwen3-0.6B [Yang et al., 2025], (iii) Gemma-3-12B-It/Gemma-3-4B-It [Team et al., 2025]. We evaluate on four benchmark datasets: GSM8K [Cobbe et al., 2021] for math word problems, CNNDM [See et al., 2017] for news summarization, XSum [Narayan et al., 2018] for extreme summarization, and MBPP [Austin et al., 2021] for Python programming.

Choice of Baselines. We compare DIVERSED and static ensemble against speculative decoding [Leviathan et al., 2023, Chen et al., 2023], lossy speculative decoding [Leviathan et al., 2023], and SpecCascade [Narasimhan et al., 2025]. Please refer to Appendix C for a more detailed experiment setting.

6.2 Experimental result and analysis

Benchmark wall-clock time and acceptance rate. Figure 4 reports the normalized wall-clock time needed for different model pairs to complete generation vs the acceptance rate in speculative decoding. We measure wall-clock time per model pair as the time required to generate a fixed number of tokens: 384 tokens for GSM8K, 128 for CNNDM and XSum, and 512 for MBPP. We find that the wall-clock time decreases monotonically as the acceptance rate increases.

Benchmark acceptance rate and generation quality. In the main Table 1, we report the acceptance rate along with the generation quality across different model pairs and various tasks. For the baseline



(a) Acceptance rate decreases approximately linearly as ensemble weight increases. Results are reported on the CNNDM dataset with target/draft model pair: Llama-3.1-8B-Instruct/Llama-3.2-1B-Instruct, and temperature is set as 0.

(b) Average number of tokens accepted per verification round (temperature = 1, draft length $N = 5$) across three target/draft model pairs. Our DIVERSED consistently surpasses all baselines by achieving a higher number of accepted tokens.

s, lossy speculative decoding, SpecCascade, and static ensemble verification are reported such that the generation quality matches the target model quality.

Notably, DIVERSED consistently achieves a larger acceptance rate than lossy speculative, SpecCascade, and Static ensemble. On the Llama-3.1-8B/Llama-3.2-1B pair, DIVERSED consistently improves the acceptance rate by at least 28% across all tasks while matching the target model generation quality. On the CNNDM dataset, DIVERSED improves the acceptance rate from 21.60% to 69.96% while also improving the generation quality. This is because the draft model also generates high-quality completions for the summarization, and DIVERSED could effectively accept those high-quality draft tokens.

Benchmark speedup: average number of tokens generated per target call. In Figure 5b, we report the average number of tokens accepted per verification round for different target/draft pairs, averaged over all datasets and generations.

Notably, DIVERSED achieves a higher average number of accepted tokens than all baselines, because it adaptively blends the target and draft distributions at each step—enlarging the acceptance region for high-confidence draft tokens while tightly controlling distributional deviation.

Figure 5a shows that the acceptance rate decreases as the ensemble weight w increases, and as $w \rightarrow 1$, the static ensemble verification becomes exactly the lossless speculative decoding. It also confirms the theoretical prediction of Lemma 3.2: the acceptance rate decreases as the ensemble weight increases across different draft lengths N .

In Figure 2, we observe a clear trade-off between summarization quality (ROUGE-2) [Lin, 2004] and inference time: the faster inference time implies the worse output quality in terms of ROUGE-2 score. Static ensemble provides a simple and controllable mechanism to traverse this Pareto front, while our proposed DIVERSED method goes beyond the Pareto front obtained by static ensemble.

7 Conclusion

We presented DIVERSED, a relaxed verification framework for speculative decoding that raises acceptance while preserving quality. Our study introduced two concrete methods: a training-free *static ensemble* verification that exactly traverses the acceptance–quality Pareto front, and DIVERSED, a learned *dynamic ensemble* that adapts per step to accept high-confidence draft tokens and control deviation from the target. Theoretically, we derived an exact expression for the expected accepted length without i.i.d. assumptions and proved that static ensembling attains the Pareto frontier. Empirically, we show that higher acceptance reliably translates into lower latency, and DIVERSED consistently achieved larger acceptance at comparable or better task quality.

Future work includes extending relaxed verification to block-level verification and exploring cross-task transfer of the dynamic head.

References

- J. Austin, A. Odena, M. Nye, M. Bosma, H. Michalewski, D. Dohan, E. Jiang, C. Cai, M. Terry, Q. Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.
- G. Bachmann, S. Anagnostidis, A. Pumarola, M. Georgopoulos, A. Sanakoyeu, Y. Du, E. Schönfeld, A. K. Thabet, and J. Kohler. Judge decoding: Faster speculative sampling requires going beyond model alignment. In *ICLR*. OpenReview.net, 2025.
- T. Cai, Y. Li, Z. Geng, H. Peng, J. D. Lee, D. Chen, and T. Dao. Medusa: Simple llm inference acceleration framework with multiple decoding heads. In *ICML*, volume 235, pages 5209–5235, 2024.
- C. Chen, S. Borgeaud, G. Irving, J.-B. Lespiau, L. Sifre, and J. Jumper. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*, 2023.
- K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- T. Dao, D. Fu, S. Ermon, A. Rudra, and C. Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *NeurIPS*, 35:16344–16359, 2022.
- T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *NeurIPS*, 35:30318–30332, 2022.
- A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.
- M. Elhoushi, A. Shrivastava, D. Liskovich, B. Hosmer, B. Wasti, L. Lai, A. Mahmoud, B. Acun, S. Agarwal, A. Roman, A. A. Aly, B. Chen, and C. Wu. Layerskip: Enabling early exit inference and self-speculative decoding. In *ACL (1)*, pages 12622–12642. Association for Computational Linguistics, 2024.
- R. Goel, M. Gagrani, W. Jeon, J. Park, M. Lee, and C. Lott. Direct alignment of draft model for speculative decoding with chat-fine-tuned llms. *arXiv preprint arXiv:2403.00858*, 2024.
- J. Hu. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv preprint arXiv:2501.03262*, 2025.
- K. Huang, X. Guo, and M. Wang. Specdec++: Boosting speculative decoding via adaptive candidate lengths. *arXiv preprint arXiv:2405.19715*, 2024.
- J. Jiang, F. Wang, J. Shen, S. Kim, and S. Kim. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*, 2024.
- T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *NeurIPS*, 35:22199–22213, 2022.
- Y. Leviathan, M. Kalman, and Y. Matias. Fast inference from transformers via speculative decoding. In *ICML*, pages 19274–19286. PMLR, 2023.
- M. Li, X. Chen, A. Holtzman, B. Chen, J. Lin, S. Yih, and V. Lin. Nearest neighbor speculative decoding for LLM generation and attribution. In *NeurIPS*, 2024a.
- Y. Li, F. Wei, C. Zhang, and H. Zhang. EAGLE: speculative sampling requires rethinking feature uncertainty. In *ICML*. OpenReview.net, 2024b.
- C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- X. Liu, L. Hu, P. Bailis, A. Cheung, Z. Deng, I. Stoica, and H. Zhang. Online speculative decoding. *arXiv preprint arXiv:2310.07177*, 2023.

- J. Mamou, O. Pereg, D. Korat, M. Berchansky, N. Timor, M. Wasserblat, and R. Schwartz. Dynamic speculation lookahead accelerates speculative decoding of large language models. In *NeurIPS Efficient Natural Language and Speech Processing Workshop*, pages 456–467. PMLR, 2024.
- X. Miao, G. Oliaro, Z. Zhang, X. Cheng, Z. Wang, Z. Zhang, R. Y. Y. Wong, A. Zhu, L. Yang, X. Shi, C. Shi, Z. Chen, D. Arfeen, R. Abhyankar, and Z. Jia. Specinfer: Accelerating large language model serving with tree-based speculative inference and verification. In *ASPLOS*, page 932–949, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703867.
- H. Narasimhan, W. Jitkrittum, A. S. Rawat, S. Kim, N. Gupta, A. K. Menon, and S. Kumar. Faster cascades via speculative decoding. In *ICLR*. OpenReview.net, 2025.
- S. Narayan, S. B. Cohen, and M. Lapata. Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745, 2018.
- Y. Ni, C. Liu, Y. Tang, K. Han, and Y. Wang. EMS-SD: Efficient multi-sample speculative decoding for accelerating large language models. In *NAACL*, pages 9307–9320. Association for Computational Linguistics, 2025.
- Y. Qin, S. Liang, Y. Ye, K. Zhu, L. Yan, Y. Lu, Y. Lin, X. Cong, X. Tang, B. Qian, S. Zhao, L. Hong, R. Tian, R. Xie, J. Zhou, M. Gerstein, D. Li, Z. Liu, and M. Sun. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *ICLR*. OpenReview.net, 2024.
- B. Rozière, J. Gehring, F. Gloeckle, S. Sootla, I. Gat, X. E. Tan, Y. Adi, J. Liu, R. Sauvestre, T. Remez, J. Rapin, A. Kozhevnikov, I. Evtimov, J. Bitton, M. Bhatt, C. C. Ferrer, A. Grattafiori, W. Xiong, A. Défossez, J. Copet, F. Azhar, H. Touvron, L. Martin, N. Usunier, T. Scialom, and G. Synnaeve. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.
- T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom. Toolformer: Language models can teach themselves to use tools. *NeurIPS*, 36: 68539–68551, 2023.
- A. See, P. J. Liu, and C. D. Manning. Get to the point: Summarization with pointer-generator networks. In *ACL (1)*, pages 1073–1083. Association for Computational Linguistics, 2017.
- G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.
- J. Wei, X. Wang, D. Schuurmans, M. Bosma, brian ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837, 2022.
- M. Yan, S. Agarwal, and S. Venkataraman. Decoding speculative decoding. In *NAACL (Long Papers)*, pages 6460–6473. Association for Computational Linguistics, 2025.
- A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- M. Yin, M. Chen, K. Huang, and M. Wang. A theoretical perspective for speculative decoding algorithm. In *NeurIPS*, volume 37, pages 128082–128117, 2024.
- H. Zhang and H. Zhao. Minimum divergence vs. maximum margin: an empirical comparison on seq2seq models. In *ICLR*, 2018.
- J. Zhang, J. Wang, H. Li, L. Shou, K. Chen, G. Chen, and S. Mehrotra. Draft& verify: Lossless large language model acceleration via self-speculative decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11263–11282, 2024.
- Y. Zhou, K. Lyu, A. S. Rawat, A. K. Menon, A. Rostamizadeh, S. Kumar, J. Kagy, and R. Agarwal. Distillspec: Improving speculative decoding via knowledge distillation. In *ICLR*. OpenReview.net, 2024.

Table 2: List of symbols and their definitions.

Symbol	Definition
N	Draft sequence length
x_0	Input prompt
$x_{1:N}$	Sampled accepted token sequence
T_N	Number of rejections when sampling a sequence of length N
M_q	Draft (small) model
M_p	Target (large) model
$q(\cdot \mid x_{1:t-1})$ or $q_t(\cdot)$	Draft (small) model sampling distribution at time t .
$p(\cdot \mid x_{1:t-1})$ or $p_t(\cdot)$	Target (small) model sampling distribution at time t .
ν	Proposed static or dynamic ensemble for between draft model and target model.

A Theoretical Justification of DIVERSED method

A.1 Proof of Theorem 4.2

In this section, we provide a detailed proof of Theorem 4.2. Let A_i denote the random variable indicating whether the i -th token is rejected. Specifically, $A_i = 1$ represents the event that the i -th draft token x_i is rejected, and the probability of this event is given by $\mathbb{P}(A_i = 1)$.

We start by computing the conditional expectation, given a draft generation $x_{1:N}$:

$$\mathbb{E}[T_N \mid \tilde{x}_{1:N}] = \sum_{t=1}^N \mathbb{P}(T_N > t \mid \tilde{x}_{1:N}) = \sum_{t=1}^N \prod_{i=1}^t \mathbb{P}(A_i = 1 \mid \tilde{x}_{1:N}) = \sum_{t=1}^N \prod_{i=1}^t \min \left\{ 1, \frac{p(x_i \mid x_{0:i-1})}{q(x_i \mid x_{0:i-1})} \right\} \quad (6)$$

Then, by the tower law, we can have the following:

$$\begin{aligned} \mathbb{E}[T_N] &= \mathbb{E}_{x_{1:N} \sim q} [\mathbb{E}[T_N \mid x_{1:N}]] = \mathbb{E}_{x_{1:N} \sim q} \left[\sum_{t=1}^N \prod_{i=1}^t \min \left\{ 1, \frac{p(x_i \mid x_{0:i-1})}{q(x_i \mid x_{0:i-1})} \right\} \right] \\ &= \sum_{x_{1:N}} \left[\sum_{t=1}^N \prod_{i=1}^t \min \left\{ 1, \frac{p(x_i \mid x_{0:i-1})}{q(x_i \mid x_{0:i-1})} \right\} \right] q(x_{1:N}) \\ &= \sum_{x_{1:N}} \left[\sum_{t=1}^N \left[\prod_{i=1}^t \min \left\{ 1, \frac{p(x_i \mid x_{0:i-1})}{q(x_i \mid x_{0:i-1})} \right\} \right] \prod_{i=1}^N q(x_i \mid x_{0:i-1}) \right] \\ &= \sum_{x_{1:N}} \left[\sum_{t=1}^N \prod_{i=1}^t \min \{ q(x_i \mid x_{0:i-1}), p(x_i \mid x_{0:i-1}) \} \prod_{i=t+1}^N q(x_i \mid x_{0:i-1}) \right] \\ &= \sum_{t=1}^N \left[\sum_{x_{1:N}} \prod_{i=1}^t \min \{ q(x_i \mid x_{0:i-1}), p(x_i \mid x_{0:i-1}) \} \prod_{i=t+1}^N q(x_i \mid x_{0:i-1}) \right] \\ &= \sum_{t=1}^N \sum_{x_{1:t}} \prod_{i=1}^t \min \{ q(x_i \mid x_{0:i-1}), p(x_i \mid x_{0:i-1}) \}. \quad (7) \end{aligned}$$

For any $t > 1$, one can use the induction rule and derive the following equation:

$$\sum_{x_{1:t}} \prod_{i=1}^t \min\{q(x_i | x_{0:i-1}), p(x_i | x_{0:i-1})\} \quad (8)$$

$$= \sum_{x_{1:t-1}} \sum_{x_t} \min\{q(x_t | x_{0:t-1}), p(x_t | x_{0:t-1})\} \prod_{i=1}^{t-1} \min\{q(x_i | x_{0:i-1}), p(x_i | x_{0:i-1})\} \quad (9)$$

$$= \sum_{x_{1:t-1}} (1 - \text{TV}(q(x | x_{0:t-1}), p(x | x_{0:t-1}))) \prod_{i=1}^{t-1} \min\{q(x_i | x_{0:i-1}), p(x_i | x_{0:i-1})\} \quad (10)$$

$$= \sum_{x_{1:t-1}} \prod_{i=t-1}^t (1 - \text{TV}(q(x | x_{0:i}), p(x | x_{0:i}))) \frac{\min\{q(x_{t-1} | x_{0:t-2}), p(x_{t-2} | x_{0:t-2})\}}{1 - \text{TV}(q(x | x_{0:t-2}), p(x | x_{0:t-2}))} \cdot \prod_{i=1}^{t-2} \min\{q(x_i | x_{0:i-1}), p(x_i | x_{0:i-1})\} \quad (11)$$

$$= \sum_{x_{1:t-1}} \prod_{i=t-1}^t (1 - \text{TV}(q(x | x_{0:i}), p(x | x_{0:i}))) \tilde{m}(x_{t-1} | x_{0:t-2}) \prod_{i=1}^{t-2} \min\{q(x_i | x_{0:i-1}), p(x_i | x_{0:i-1})\} \quad (12)$$

$$= \sum_{x_{1:t-1}} \prod_{i=0}^t (1 - \text{TV}(q(x | x_{0:i}), p(x | x_{0:i}))) \prod_{i=1}^{t-1} \tilde{m}(x_i | x_{0:i-1}) \quad (13)$$

$$= \mathbb{E}_{x_{1:t-1} \sim \tilde{m}} \left[\prod_{i=0}^t (1 - \text{TV}(q(x | x_{0:i}), p(x | x_{0:i}))) \right] \quad (14)$$

Combined with Equation 7, we finish the proof.

A.2 Proof of Lemma 3.2

Proof. The expected probability of accepting the $\tilde{x}_t$ given the $x_{0:t-1}$ is $\mathbb{E}_{x \sim q_t} \left[\min \left\{ 1, \frac{\nu_t(x)}{q_t(x)} \right\} \right]$ and can be shown as:

$$\begin{aligned} \mathbb{E}_{x \sim q_t} \left[\min \left\{ 1, \frac{\nu_t(x)}{q_t(x)} \right\} \right] &= \sum_x \min \left\{ 1, \frac{\nu_t(x)}{q_t(x)} \right\} q_t(x) = \sum_x \min \{q_t(x), \nu_t(x)\} \\ &= \sum_x \min \{q_t(x), wp_t(x) + (1-w)q_t(x)\} \\ &= \sum_x q_t(x) + \sum_x \min \{0, w(p_t(x) - q_t(x))\} \\ &= 1 - w + \sum_x w \min \{q_t(x), p_t(x)\} \\ &= 1 - w + w(1 - \text{TV}(q_t(x), p_t(x))) \\ &= 1 - w \text{TV}(q_t(x), p_t(x)) \end{aligned}$$

where $\nu_t(x) = wp_t(x) + (1-w)q_t(x)$ as defined in Equation (2). $\square$

A.3 Proof of Lemma 4.1

Proof. Applying the Bayes rule, one can have:

$$\begin{aligned}
\mathbb{P}(x_t = x \mid x_t \text{ is accepted}) &= \frac{\mathbb{P}(x_t \text{ is accepted}, x_t = x)}{\mathbb{P}(x_t \text{ is accepted})} = \frac{\mathbb{P}(x_t \text{ is accepted} \mid x_t = x) \mathbb{P}(x_t = x)}{\sum_x \mathbb{P}(x_t \text{ is accepted}, x_t = x)} \\
&= \frac{\min\{1, \frac{p(x \mid x_{0:t-1})}{q(x \mid x_{0:t-1})}\} q(x \mid x_{0:t-1})}{\sum_x \min\{1, \frac{p(x \mid x_{0:t-1})}{q(x \mid x_{0:t-1})}\} q(x \mid x_{0:t-1})} \\
&= \frac{\min\{q(x \mid x_{0:t-1}), p(x \mid x_{0:t-1})\}}{\sum_x \min\{q(x \mid x_{0:t-1}), p(x \mid x_{0:t-1})\}} \\
&= \frac{\min\{q(x \mid x_{0:t-1}), p(x \mid x_{0:t-1})\}}{1 - \text{TV}(q(x \mid x_{0:t-1}), p(x \mid x_{0:t-1}))}
\end{aligned}$$

□

A.4 Proof of Proposition 3.1

Proof.

$$\begin{aligned}
\mathbb{P}_{\nu_t}(\text{reject}) &= 1 - \sum_x \min\left\{1, \frac{wp(x) + (1-w)q(x)}{q(x)}\right\} q(x) = 1 - \sum_x \min\{q(x), wp(x) + (1-w)q(x)\} \\
&= w - \sum_x \min\{wq(x), wp(x)\} \\
&= w(1 - \sum_x \min\{q(x), p(x)\}) \\
&= w\text{TV}(p, q).
\end{aligned}$$

$$\begin{aligned}
\text{TV}(\nu_t, p) &= \sum_x \max\{0, wp(x) + (1-w)q(x) - p(x)\} = \sum_x \max\{0, (1-w)(q(x) - p(x))\} \\
&= (1-w)\text{TV}(q, p) \\
&= (1-w)\text{TV}(p, q).
\end{aligned}$$

Thus, we can have:

$$\mathbb{P}_{\nu_t}(\text{reject}) + \text{TV}(\nu_t, p) = \text{TV}(q, p). \quad (15)$$

□

B Implementation Detail of DIVERSED

The neural network architecture of the proposed DIVERSED consists of a single fully connected layer, whose input and output dimensions match the input vector. The activation function used is the sigmoid function.

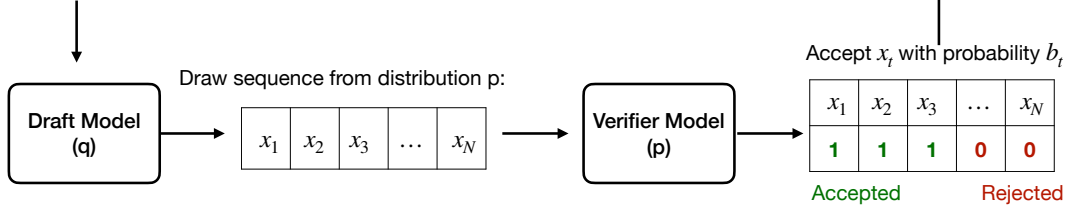


Figure 6: Overview of the speculative decoding framework. Our DIVERSED method employs different acceptance rules b_t (described in Section 3) to accept or reject the sampled sequences.

Algorithm 2 Inference pipeline of DIVERSED.

Inputs: Horizon T ; draft length N ; prompt x_0 ; draft model $\{q_t\}$; target model $\{p_t\}$.

Outputs: Sequence $x_{1:T}$.

```

1:  $n \leftarrow 1$ ;
2: while  $n \leq T$  do
3:   for  $t = n, \dots, n + N$  do
4:     sample  $\tilde{x}_t \sim q_t(\cdot \mid x_{0:n-1}, \tilde{x}_{n:t-1})$  ▷ draft rollout
5:     cache hidden states  $h_{n:n+N}^q$  and distribution  $q_{n:n+N}$ .
6:   for  $t = n, \dots, n + N$  do
7:     compute  $p_t(\cdot \mid x_{0:n-1}, \tilde{x}_{n:t-1})$ , and cache hidden states  $h_{n:n+N}^p$  ▷ target parallel verification
8:     Compute the relaxed accept rule  $b_{\nu_t^\theta}$  and  $P_{\nu_t^\theta}$  in equation (5) ▷ Dynamic Ensemble
9:   for  $t = n$  to  $n + N$  do
10:    Draw  $u \sim \text{Uniform}(0, 1)$ .
11:    if  $u \leq b_{\nu_t^\theta}(\tilde{x}_t \mid x_{1:n-1}, \tilde{x}_{n:t-1})$  then
12:      accept  $\tilde{x}_t$ ;
13:       $x_n \leftarrow \tilde{x}_t$ ;
14:       $n \leftarrow n + 1$ 
15:    else
16:      reject;
17:      sample  $x_n \sim P_{\nu_t^\theta}$ ;
18:       $n \leftarrow n + 1$ ; break ▷ discard  $\tilde{x}_{t+1:T}$  and resume

```

C Extended Experimental Analysis

C.1 Experimental Settings

We compare DIVERSED and static ensemble verification against speculative decoding, lossy speculative decoding [Leviathan et al., 2023], and SpecCascade [Narasimhan et al., 2025] on three target/draft model pairs: (i) Llama-3.1-8B-Instruct/Llama-3.2-1B-Instruct [Dubey et al., 2024], (ii) Qwen3-8B/Qwen3-0.6B [Yang et al., 2025], (iii) Gemma-3-12B-It/Gemma-3-4B-It [Team et al., 2025]. We evaluate on four benchmarks: GSM8K [Cobbe et al., 2021](math word problems), CNNDM [See et al., 2017](news summarization), XSum [Narayan et al., 2018](extreme summarization), and MBPP [Austin et al., 2021](Python programming).

C.2 Implementation Details

We train separate ensembles per task. For GSM8K and MBPP, the outcome reward is task accuracy; for CNNDM and XSum, we use ROUGE-2 [Lin, 2004] following Zhang and Zhao [2018]. We test temperatures 0 (greedy) and 1, and draft lengths $N \in \{3, 5, 7\}$. Generation budgets are 128 tokens for CNNDM and XSum, 384 for GSM8K, and 512 for MBPP. We use few-shot prompting by supplying input-output exemplars that enable the LLM to infer the desired generation style. All training and inference run on $8 \times$ A100-40GB GPUs.

Table 3: Hyper-parameter configuration for each benchmark dataset.

	CNNDM	XSUM	MBPP	GSM8K
Generation budget	128	128	384	512
Hyper Parameter λ	0.5	0.5	0.3	0.1
Draft length N		{3, 5, 7}		
Rollout Number		4		
Learning Rate		5e-4		
Training Epochs		2		
(a) Target/Draft model pair is Llama3.1-8B-Instruct/Llama3.2-1B-Instruct.				
Generation budget	128	128	384	512
Hyper Parameter λ	0.5	0.5	0.3	0.3
Draft length N		{3, 5, 7}		
Rollout Number		4		
Learning Rate		5e-4		
Training Epochs		2		
(b) Target/Draft model pair is Qwen3-8B/Qwen3-0.6B.				
Generation budget	128	128	384	512
Hyper Parameter λ	0.5	0.5	0.3	0.35
Draft length N		{3, 5, 7}		
Rollout Number		4		
Learning Rate		5e-4		
Training Epochs		2		
Training and testing hardware	8× A100-40GB GPUs			

(c) Target/Draft model pair is gemma-3-12b-it/gemma-3-4b-it.

C.3 More Experiment Results

Figure 7 illustrates Lemma 4.1 with Gaussian distributions: accepted tokens follow the min-kernel, whereas the combination of accepted and resampled tokens recovers the target distribution. To achieve large speedups, the acceptance proportion must be high [Yan et al., 2025, Huang et al., 2024, Narasimhan et al., 2025]. This suggests an alternative alignment method that could potentially improve the generation quality by closing the gap between the min-kernel and the target distribution.

Table 4: Reported acceptance rate and generation quality. Columns are grouped by task. Each block compares three baselines and our method for a given *Target/Draft* pair. All results use temperature 0 and draft length $N = 5$. Our DIVERSED enjoys a higher acceptance rate and also maintains comparable quality to baselines.

	GSM8K		CNNDM		Xsum		MBPP	
	Accept Rate ($\uparrow$)	Quality Acc ($\uparrow$)	Accept Rate ($\uparrow$)	Quality ROUGE-2 ($\uparrow$)	Accept Rate ($\uparrow$)	Quality ROUGE-2 ($\uparrow$)	Accept Rate ($\uparrow$)	Quality pass@1 ($\uparrow$)
Autoregressive	NA	80%	NA	11.29	NA	8.42	NA	62%
SD	61.53%	80%	71.70%	11.26	50.24%	8.36	72.18%	62%
Static Ensemble	82.58%	79%	85.04%	11.45	61.44%	7.53	84.67%	61%
DIVERSED (ours)	84.82%	80%	81.92%	12.37	94.81%	7.66	85.03%	62%
(a) Target/Draft model pair is Llama-3.1-8B/Llama-3.2-1B.								
Autoregressive	NA	92%	NA	11.48	NA	5.38	NA	57%
SD	63.70%	92%	36.40%	11.38	19.56%	5.35	70.27%	57%
Static Ensemble	67.58%	92%	41.60%	11.85	34.14%	5.39	75.70%	57%
DIVERSED (ours)	69.68%	93%	58.46%	12.64	38.12%	5.54	74.83%	59%
(b) Target/Draft model pair is Qwen3-8B/Qwen3-0.6B.								
Autoregressive	NA	93%	NA	10.87	NA	9.16	NA	66%
SD	84.15%	92%	40.39%	10.62	35.54%	9.04	83.89%	66%
Static Ensemble	91.66%	92%	54.48%	10.84	56.31%	7.90	86.69%	65%
DIVERSED (ours)	92.35%	93%	66.39%	11.47	63.38%	7.22	90.75%	64%
(c) Target/Draft model pair is Gemma-3-12B/Gemma-3-4B.								

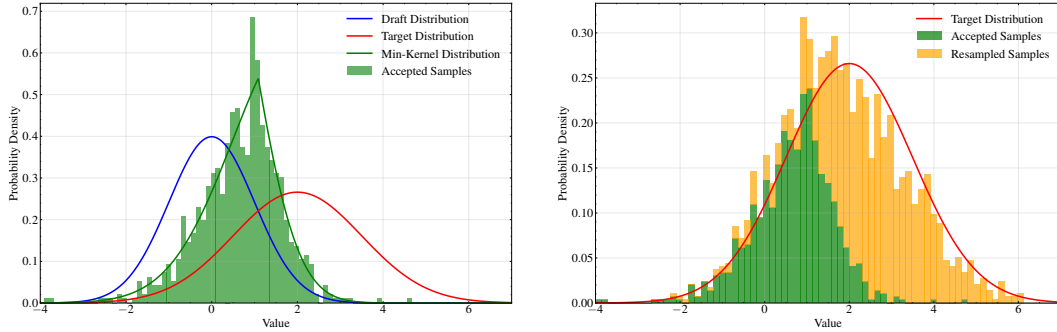


Figure 7: Illustration of speculative decoding with Gaussian draft and target distributions. **(Left):** The accepted sample distribution follows the min-kernel distribution. **(Right):** The combined accepted and resampled distribution follows the target distribution.

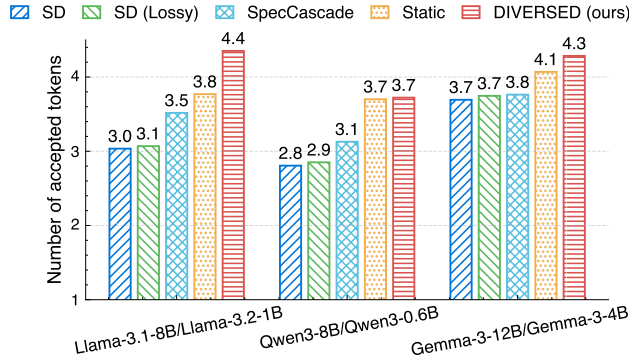


Figure 8: Average number of tokens accepted per verification round (temperature = 0.1, draft length $N = 5$) across three target/draft model pairs. Our DIVERSED consistently surpasses all baselines by achieving a higher number of accepted tokens.

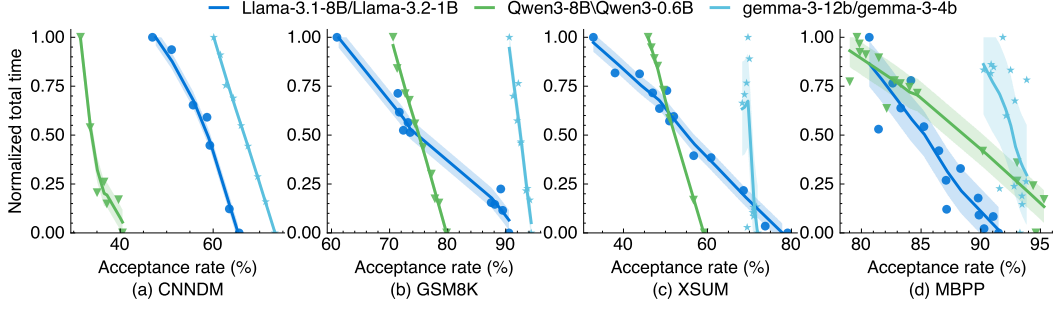


Figure 9: Results on the CNNDM, XSUM, MBPP, and GSM8k datasets for three target/draft pairs: normalized total time (lower is better) versus acceptance rate (higher is better) under speculative decoding with $N=3$. Model pairs with higher acceptance achieve lower decoding time.

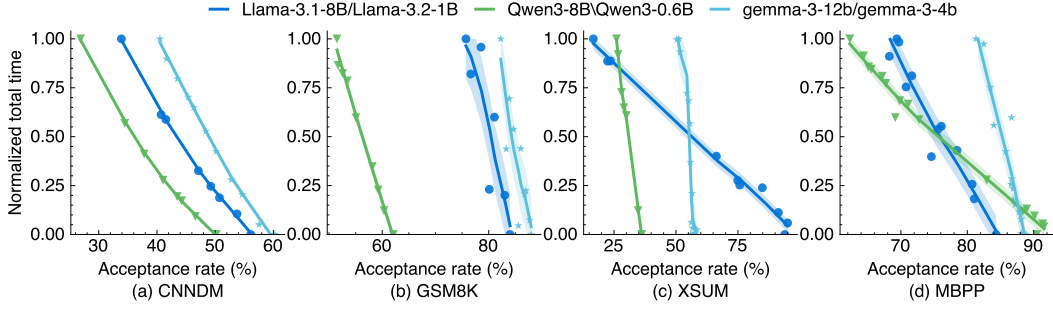


Figure 10: Results on the CNNDM, XSUM, MBPP, and GSM8k datasets for three target/draft pairs: normalized total time (lower is better) versus acceptance rate (higher is better) under speculative decoding with $N=7$. Model pairs with higher acceptance achieve lower decoding time.

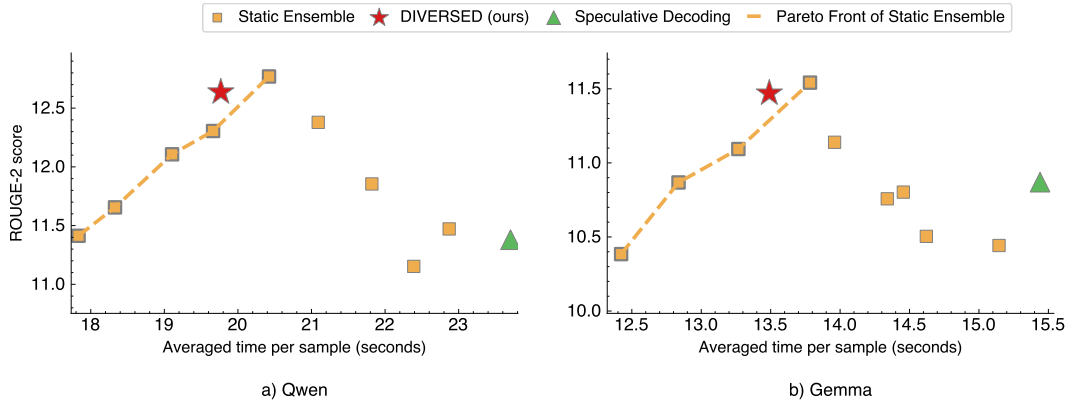


Figure 11: Our DIVERSED outperforms the Pareto front of the static ensemble and other baselines. Results are reported on the CNNDM dataset using the target/draft model pair: Qwen3-8B/Qwen3-0.6B and Gemma-3-12b-it/Gemma-3-4b-it, with temperature set to 0.

Table 5: Reported acceptance rate and generation quality, under different temperature. Columns are grouped by task. Each block compares three baselines and our method for a given *Target/Draft* pair. All results use a temperature of 0.1 and a draft length $N = 5$. Our DIVERSED enjoys a higher acceptance rate and also maintains comparable quality to baselines.

Method	GSM8K		CNNDM		XSum		MBPP	
	Accept Rate (↑)	Quality Acc (↑)	Accept Rate(↑)	Quality ROUGE-2 (↑)	Accept Rate(↑)	Quality ROUGE-2 (↑)	Accept Rate(↑)	Quality pass@1 (↑)
Autoregressive	NA	76%	NA	9.94	NA	7.56	NA	62%
SD	66.41%	76%	17.69%	9.92	37.67%	7.82	72.94%	61%
SD (Lossy)	67.42%	76%	17.85%	10.17	38.83%	7.97	73.37%	61%
SpecCascade	68.01%	76%	52.20%	10.19	40.22%	7.53	84.15%	61%
Static Ensemble	72.36%	75%	87.58%	10.46	51.20%	7.31	80.66%	61%
DIVERSED (ours)	89.14%	76%	92.11%	11.65	64.75%	7.52	89.78%	62%
(a) Target/Draft model pair is Llama-3.1-8B/Llama-3.2-1B.								
Autoregressive	NA	92%	NA	11.02	NA	5.21	NA	57%
SD	58.83%	92%	34.19%	10.35	16.45%	5.26	69.49%	56%
SD (lossy)	59.85%	92%	35.32%	10.74	19.56%	4.90	69.71%	56%
SpecCascade	59.71%	92%	37.67%	11.30	32.98%	4.91	78.32%	56%
Static Ensemble	88.75%	92%	40.67%	10.90	30.80%	5.19	79.86%	55%
DIVERSED (ours)	91.84%	92%	46.40%	11.37	45.87%	5.25	75.56%	59%
(b) Target/Draft model pair is Qwen3-8B/Qwen3-0.6B.								
Autoregressive	NA	93%	NA	9.01	NA	8.27	NA	66%
SD	81.85%	92%	44.27%	9.00	39.63%	8.27	83.85%	67%
SD (Lossy)	82.75%	92%	48.73%	10.71	40.08%	8.37	84.31%	66%
SpecCascade	83.50%	93%	48.72%	10.72	39.42%	8.79	84.68%	67%
Static Ensemble	89.67%	91%	47.47%	10.74	62.65%	8.31	88.34%	66%
DIVERSED (ours)	90.26%	92%	64.26%	11.60	71.22%	7.84	91.23%	67%
(c) Target/Draft model pair is Gemma-3-12B/Gemma-3-4B.								

C.4 GSM8K Case Study

Case Study on the GSM8K dataset using Llama model.

Question: Marissa is hiking a 12-mile trail. She took 1 hour to walk the first 4 miles, then another hour to walk the next two miles. If she wants her average speed to be 4 miles per hour, what speed (in miles per hour) does she need to walk the remaining distance?

Autoregressive	Marissa walks the remaining 6 miles at her desired average speed of 4 miles per hour, so the time it will take her to walk the remaining distance is $6 / 4 = 1.5$ hours. So in total she will have spent $1.5 + 1 + 1 = 3.5$ hours walking the 12-mile trail. ###6
SD	She has walked $4 + 2 = 6$ miles in 2 hours. So, she has $12 - 6 = 6$ miles left to walk. She wants to walk the entire trail in $12 / 4 = 3$ hours. So, she needs to walk the remaining 6 miles in $3 - 2 = 1$ hour. Therefore, she needs to walk the remaining 6 miles in $6 / 1 = 6$ miles per hour. ###6
SD (Lossy)	She has walked $4 + 2 = 6$ miles in 2 hours. So, she has $12 - 6 = 6$ miles left to walk. She wants her average speed to be 4 miles per hour, so she needs to walk the remaining 6 miles in $6 / 4 = 1.5$ hours. Therefore, she needs to walk the remaining distance at a speed of $6 / 1.5 = 4$ miles per hour. ###4
SpecCascade	First, let's find out how long it will take her to walk the entire trail. We know that she will need to walk 12 miles at a speed of 4 miles per hour, so the time it will take her to walk the entire trail is $12 / 4 = 3$ hours. She already walked 6 miles and it took her 2 hours to do so, so the time she has left is $3 - 2 = 1$ hour. Therefore, she needs to walk $12 - 6 = 6$ miles in 1 hour, so she needs to walk at a speed of $6 / 1 = 6$ miles per hour. ###6
Static Ensemble	She has walked 4 miles in 1 hour, and 2 miles in another hour, so the total distance covered is 6 miles in 2 hours, which means her average speed is $6 / 2 = 3$ miles per hour. To cover the remaining 6 miles, she needs to cover 3 miles per hour, which means 6 miles in 2 hours. Therefore, her average speed must be $6 / 2 = 3$ miles per hour. ###3.
DIVERSED	She has walked $4 + 2 = 6$ miles in $1 + 1 = 2$ hours. She has $12 - 6 = 6$ miles left to walk. She wants to walk the entire trail in $12 / 4 = 3$ hours. So she has $3 - 2 = 1$ hour to walk the remaining 6 miles. Therefore, she needs to walk $6 / 1 = 6$ miles per hour. ###6

Case Study on the GSM8K dataset using Qwen model.

Question: Marissa is hiking a 12-mile trail. She took 1 hour to walk the first 4 miles, then another hour to walk the next two miles. If she wants her average speed to be 4 miles per hour, what speed (in miles per hour) does she need to walk the remaining distance?

Autoregressive	To calculate the total time she needs to complete the trail for an average speed of 4 mph, we calculate the total time as total distance divided by average speed, which is $12 \text{ miles} / 4 \text{ mph} = 3 \text{ hours}$. So far, she has spent $1 + 1 = 2 \text{ hours}$. This leaves her with $3 - 2 = 1 \text{ hour}$ to walk the remaining distance. She has $12 - 4 - 2 = 6 \text{ miles}$ left. Therefore, she needs to walk 6 miles in 1 hour, which is a speed of 6 mph. ###6
SD	The total distance of the trail is 12 miles. So far, she has walked $4 + 2 = 6 \text{ miles}$. Therefore, she has $12 - 6 = 6 \text{ miles}$ left. She wants her average speed to be 4 miles per hour for the entire trail. To find the total time she has for the whole trail, we can use the formula: $\text{total time} = \text{total distance} / \text{average speed} = 12 \text{ miles} / 4 \text{ mph} = 3 \text{ hours}$. She has already spent $1 + 1 = 2 \text{ hours}$. Therefore, she has $3 - 2 = 1 \text{ hour}$ left to walk the remaining 6 miles. To find the required speed, we can use the formula: $\text{speed} = \text{distance} / \text{time} = 6 \text{ miles} / 1 \text{ hour} = 6 \text{ miles per hour}$. So, she needs to walk the remaining distance at 6 miles per hour. ###6
SD (Lossy)	To find the average speed, we need to divide the total distance by the total time. Marissa has already walked $4 + 2 = 6 \text{ miles}$. She has $12 - 6 = 6 \text{ miles}$ left. She has already spent $1 + 1 = 2 \text{ hours}$. To have an average speed of 4 miles per hour, she needs to complete the 12-mile trail in $12 / 4 = 3 \text{ hours}$. She has $3 - 2 = 1 \text{ hour}$ left to walk the remaining 6 miles. Therefore, she needs to walk at a speed of 6. ###6
SpecCascade	To find the average speed, we need to divide the total distance by the total time. Marissa has already walked $4 + 2 = 6 \text{ miles}$. She has $12 - 6 = 6 \text{ miles}$ left. She has already spent $1 + 1 = 2 \text{ hours}$. To have an average speed of 4 miles per hour, the total time for the entire trail should be $12 / 4 = 3 \text{ hours}$. Therefore, she has $3 - 2 = 1 \text{ hour}$ left to walk the remaining 6 miles. So, she needs to walk at a speed of 6. ###6
Static Ensemble	To find the average speed, we need to divide the total distance by the total time. Marissa has already walked $4 + 2 = 6 \text{ miles}$. She has $12 - 6 = 6 \text{ miles}$ left. She has already spent $1 + 1 = 2 \text{ hours}$. She wants her average speed to be 4 miles per hour, so the total time for the entire trail should be $12 / 4 = 3 \text{ hours}$. Therefore, she has $3 - 2 = 1 \text{ hour}$ left to walk the remaining 6 miles. So, she needs to walk at a speed of $6 / 1 = 6 \text{ miles per hour}$. ###6
DIVERSED	To find the average speed, we need to divide the total distance by the total time. The total distance is 12 miles. The total time is 2 hours for the first 6 miles. She wants her average speed to be 4 miles per hour, so the total time should be $12 / 4 = 3 \text{ hours}$. Therefore, she needs to walk the remaining 6 miles in $3 - 2 = 1 \text{ hour}$. So her speed for the remaining distance should be $6 / 1 = 6 \text{ miles per hour}$. ###6

Case Study on the GSM8K dataset using Gemma model.

Question: Marissa is hiking a 12-mile trail. She took 1 hour to walk the first 4 miles, then another hour to walk the next two miles. If she wants her average speed to be 4 miles per hour, what speed (in miles per hour) does she need to walk the remaining distance?

Autoregressive	The remaining distance is $12 - 4 - 2 = 6$ miles. The total time she wants to take for the whole trail is $12 / 4 = 3$ hours. She has already spent $1 + 1 = 2$ hours hiking, so she has $3 - 2 = 1$ hour left. The required speed for the last 6 miles in 1 hour is $6 / 1 = 6$ miles per hour. ###6
SD	Total distance is 12 miles, and she wants her average speed to be 4 miles per hour. So, the total time she can spend hiking is $12 / 4 = 3$ hours. She has already spent $1 + 1 = 2$ hours. Therefore, she can hike the remaining distance for $3 - 2 = 1$ hour. The remaining distance is $12 - 4 - 2 = 6$ miles. So, she needs to walk 6 miles in 1 hour, which means she needs to walk at a speed of 6 miles per hour. ###6
SD (Lossy)	The total distance of the trail is 12 miles. Marissa has already walked $4 + 2 = 6$ miles. The remaining distance is $12 - 6 = 6$ miles. She wants her average speed to be 4 miles per hour for the entire 12-mile trail. So, the total time she should take to complete the 12-mile trail is $12 / 4 = 3$ hours. She has already taken $1 + 1 = 2$ hours. The remaining time she has is $3 - 2 = 1$ hour. To walk the remaining 6 miles in 1 hour, she needs to walk at a speed of $6 / 1 = 6$ miles per hour. ###6
SpecCascade	Total distance is 12 miles. Distance covered is $4 + 2 = 6$ miles. Remaining distance is $12 - 6 = 6$ miles. She wants her average speed to be 4 miles per hour over the entire 12-mile trail. Total time to complete the trail should be $12 / 4 = 3$ hours. Time spent so far is $1 + 1 = 2$ hours. Remaining time is $3 - 2 = 1$ hour. So she needs to walk the remaining 6 miles in 1 hour. Therefore, the required speed is $6 / 1 = 6$ miles per hour. ###6
Static Ensemble	Total distance is 12 miles. Marissa has already walked $4 + 2 = 6$ miles. The remaining distance is $12 - 6 = 6$ miles. She wants her average speed to be 4 miles per hour over the entire 12-mile trail. So the total time she should take to walk the entire trail is $12 / 4 = 3$ hours. She has already walked for $1 + 1 = 2$ hours. Therefore, she has $3 - 2 = 1$ hour to walk the remaining 6 miles. The speed she needs to walk the remaining distance is $6 \text{ miles} / 1 \text{ hour} = 6 \text{ miles per hour}$. ###6
DIVERSED	The total distance she has to cover is 12 miles. She already covered $4 + 2 = 6$ miles. She needs to cover $12 - 6 = 6$ miles. So she needs to cover the remaining 6 miles in the remaining time. To find her average speed for the entire hike, she needs to calculate the total time taken for the entire hike and then determine how much time she has left to cover the remaining 6 miles. Her average speed is 4 miles per hour for the entire hike. So her total time should be $12 \text{ miles} / 4 \text{ miles per hour} = 3 \text{ hours}$. She has already taken $1 + 1 = 2$ hours. Therefore, she has $3 - 2 = 1$ hour left to cover the remaining 6 miles. To find the speed she needs to walk the remaining 6 miles, we can calculate $6 \text{ miles} / 1 \text{ hour} = 6 \text{ miles per hour}$. ###6

C.5 CNNDM Case Study

Case Study on the CNNDM dataset using Llama model.

News: President Barack Obama held up the shrinking Florida Everglades today as proof positive that climate change is real and is threatening America's 'national treasures,' as well as the economies of their surrounding towns, which rely heavily on tourist dollars. 'Climate change can no longer be denied. It can't be edited out,' he said this afternoon in a speech at Everglades National Park commemorating Earth Day. 'It can't be omitted from the conversation. And action can no longer be delayed.' 'That's why I've committed the United States to lead the world in combatting this threat,' he said, throwing a spotlight on his administration's efforts to reduce the rate of global warming through the reduction of carbon emissions. The day trip to South Florida, on which Obama was accompanied by Bill Nye 'the Science Guy,' also highlighted the massive amount of fuel it takes to power Air Force One, however, 9,000 gallons, round trip. Scroll down for video. President Barack Obama takes a walking tour of the Anhinga Trail at Everglades National Park. Obama visited the subtropical swamps of the park today as part of a push to get Americans thinking and talking about the damage climate change is causing close to home. Obama held up the shrinking Florida Everglades today as proof positive that climate change is real and is threatening America's 'national treasures,' as well as the economies of their surrounding towns, which rely heavily on tourist dollars. 'Climate change can no longer be denied. It can't be edited out,' he said this afternoon in a speech at Everglades National Park commemorating Earth Day. 'It can't be omitted from the conversation. And action can no longer be delayed' Vehicles in the president's motorcade are seen in the Everglades National Park where Obama stopped for a visit on the Anhinga trail in Homestead. Asked earlier this week if Obama's gas-guzzling trip to Florida would undermine his Earth Day message, the White House said it would not and reaffirmed Obama's efforts to reduce pollutants. On of the helicopters escorting Marine One with President Barack Obama onboard, prepares to land at Miami's International Airport. I love the smell of jet fuel, Nye told reporters as he waited to board the president's 747 this morning for the flight to Florida, during which the White House said he would film a video with Obama. Asked earlier this week if Obama's gas-guzzling trip to Florida would undermine his Earth Day message, the White House said it would not and reaffirmed Obama's efforts to reduce pollutants. Obama took a half-hour walking tour of the the Anhinga Trail at the 1.5-million-acre national park today before making a plea for Americans to save the Everglades and other natural gems by making environmentally-conscious decisions. 'I can't think of a better way to spend Earth Day than in one of our nation's greatest natural treasures, the Everglades,' he told his audience, calling the swamp, which he acknowledged is not technically a swamp, 'magical.' 'Climate change is threatening this treasure and the communities that depend on it,' Obama warned. 'If we don't act, there may not be an Everglades as we know it.' The president said the fallout from rising temperatures is 'not a problem for another generation.' At least, 'not anymore,' he said. 'This is a problem now. It has serious implications for the way we live right now.' Obama said the rising sea level in South Florida and inward flow of salt water is evidence that climate change is already having a negative effect on communities. 'This is not some impossible problem that we cannot solve. We can solve it if we've got some political will,' he said. 'And we can solve it in a way that creates jobs. We can solve it in a way that doesn't disrupt our economy but enhances our economy. And it's a bipartisan issue.' 'That's why I've committed the United States to lead the world in combatting this threat,' he said, throwing a spotlight on his administration's efforts to reduce the rate of global warming through the reduction of carbon emissions. An alligator is spotted sunning itself along the Anhinga Trail at Everglades National Park as Obama took a tour. Wiping sweat from his brow as he braved the subtropical temperatures to make his mark on Earth Day, the president mingled with crowd-goers at the mid-day event before returning to Washington on Air Force One. Previewing the day trip, the White House had reiterated Obama's belief that 'no challenge poses a greater threat to future generations than climate change.' It's statement pointed out that last year was the warmest year on record and that 14 of the 15 of earth's hottest years happened this century. It said, 'Climate change is also affecting some of the most iconic places in our country, from disappearing glaciers in Glacier National Park to dying Joshua Trees in Joshua Tree National Park.' 'Recognized worldwide as a unique and treasured landscape, the Everglades is a perfect example of the threat we face from climate change, including rising sea levels that result in shoreline erosion and increased flooding.' Obama said it is this generation's 'turn to ensure that' natural monuments remain the 'birthright of all Americans for generations to come.' Wiping sweat from his brow as he braved the subtropical temperatures to make his mark on Earth Day, the president mingled with crowd-goers at the mid-day event before returning to Washington on Air Force One. The White House also asserted that climate change was already displacing animals and plants, including tropical orchids, some of which are only found in south Florida. 'We are blessed with the most beautiful God-given landscape in the world,' the president said today. 'It's an incredible bounty that's been given to us. But we've got to be good stewards for it. We have to take care of it.' Obama said it is this generation's 'turn to ensure that' natural monuments remain the 'birthright of all Americans for generations to come.' 'So many

people here are active in your communities, doing what's needed. The young people who are here, the next generation, they're way ahead of us in understanding how important this is. 'Let's make sure we don't disappoint them. Let's stand up and do what's right before it's too late,' he concluded. Obama was accompanied by Bill Nye 'the Science Guy' on the day trip. He also gave Florida Rep. Debbie Wasserman Schultz, who he's seen here hugging, and Interior Secretary Sally Jewell, second from left, a ride."

Autoregressive	President Barack Obama uses Florida Everglades visit to emphasize dangers of climate change. Air Force One uses 9,000 gallons of fuel, round trip, for the trip. White House says Obama's trip will not undermine his Earth Day message.
SD	President Barack Obama held up the shrinking Florida Everglades as proof of climate change. Obama said climate change is a problem now, not a problem for another generation. The president said the rising sea level in South Florida and inward flow of salt water is evidence that climate change is already having a negative effect on communities. Obama said the U.S. will lead the world in combatting climate change through the reduction of carbon emissions. Obama's trip to Florida was criticized for the amount of fuel it took to power Air Force One. Obama was accompanied by Bill Nye 'the Science Guy' on the day trip.
SD (Lossy)	President Obama visits Everglades National Park today to highlight the threat of climate change. He says the damage is real and immediate and can no longer be denied. Obama says the U.S. will lead the world in reducing carbon emissions. The president's trip to the Everglades is being criticized for the fuel it uses, 9,000 gallons, round trip. Obama says the damage from climate change is not a problem for another generation, but a problem now. The president says the U.S. can solve the problem of climate change and create jobs at the same time.
SpecCascade	President Obama visits Everglades National Park today to highlight the threat of climate change. Obama says climate change is already having a negative effect on communities in South Florida and that there's not a problem for another generation, 'not anymore.' The president calls the Everglades 'magical' and warns that the threat of climate change is not just about the environment, but also about the economies of surrounding towns that rely heavily on tourist dollars. Obama also says he wants to lead the world in combating this threat and that he's committed the United States to reducing carbon emissions.
Static Ensemble	President Barack Obama held up the shrinking Florida Everglades as proof positive that climate change is real and is threatening America's 'national treasures'. Obama said climate change can no longer be denied and action can no longer be delayed. The president said the fallout from rising temperatures is 'not a problem for another generation.' At least, 'not anymore,' he said. 'This is a problem now. It has serious implications for the way we live right now.' Obama said the rising sea level in South Florida and inward flow of salt water is evidence that climate change is already having a negative effect on communities.
DIVERSED	President Barack Obama holds up the shrinking Florida Everglades as proof that climate change is real and is threatening America's 'national treasures.' Obama says climate change can no longer be denied and that action can no longer be delayed. The president visited the subtropical swamps of the park today as part of a push to get Americans thinking and talking about the damage climate change is causing close to home. Obama's trip to Florida was accompanied by Bill Nye 'the Science Guy' and was part of a push to get Americans thinking and talking about the damage climate change is causing close to home.

News: President Barack Obama held up the shrinking Florida Everglades today as proof positive that climate change is real and is threatening America's 'national treasures,' as well as the economies of their surrounding towns, which rely heavily on tourist dollars. 'Climate change can no longer be denied. It can't be edited out,' he said this afternoon in a speech at Everglades National Park commemorating Earth Day. 'It can't be omitted from the conversation. And action can no longer be delayed.' 'That's why I've committed the United States to lead the world in combatting this threat,' he said, throwing a spotlight on his administration's efforts to reduce the rate of global warming through the reduction of carbon emissions. The day trip to South Florida, on which Obama was accompanied by Bill Nye 'the Science Guy,' also highlighted the massive amount of fuel it takes to power Air Force One, however, 9,000 gallons, round trip. Scroll down for video. President Barack Obama takes a walking tour of the Anhinga Trail at Everglades National Park. Obama visited the subtropical swamps of the park today as part of a push to get Americans thinking and talking about the damage climate change is causing close to home. Obama held up the shrinking Florida Everglades today as proof positive that climate change is real and is threatening America's 'national treasures,' as well as the economies of their surrounding towns, which rely heavily on tourist dollars. 'Climate change can no longer be denied. It can't be edited out,' he said this afternoon in a speech at Everglades National Park commemorating Earth Day. 'It can't be omitted from the conversation. And action can no longer be delayed' Vehicles in the president's motorcade are seen in the Everglades National Park where Obama stopped for a visit on the Anhinga trail in Homestead. Asked earlier this week if Obama's gas-guzzling trip to Florida would undermine his Earth Day message, the White House said it would not and reaffirmed Obama's efforts to reduce pollutants. On of the helicopters escorting Marine One with President Barack Obama onboard, prepares to land at Miami's International Airport. I love the smell of jet fuel, Nye told reporters as he waited to board the president's 747 this morning for the flight to Florida, during which the White House said he would film a video with Obama. Asked earlier this week if Obama's gas-guzzling trip to Florida would undermine his Earth Day message, the White House said it would not and reaffirmed Obama's efforts to reduce pollutants. Obama took a half-hour walking tour of the the Anhinga Trail at the 1.5-million-acre national park today before making a plea for Americans to save the Everglades and other natural gems by making environmentally-conscious decisions. 'I can't think of a better way to spend Earth Day than in one of our nation's greatest natural treasures, the Everglades,' he told his audience, calling the swamp, which he acknowledged is not technically a swamp, 'magical.' 'Climate change is threatening this treasure and the communities that depend on it,' Obama warned. 'If we don't act, there may not be an Everglades as we know it.' The president said the fallout from rising temperatures is 'not a problem for another generation.' At least, 'not anymore,' he said. 'This is a problem now. It has serious implications for the way we live right now.' Obama said the rising sea level in South Florida and inward flow of salt water is evidence that climate change is already having a negative effect on communities. 'This is not some impossible problem that we cannot solve. We can solve it if we've got some political will,' he said. 'And we can solve it in a way that creates jobs. We can solve it in a way that doesn't disrupt our economy but enhances our economy. And it's a bipartisan issue.' 'That's why I've committed the United States to lead the world in combatting this threat,' he said, throwing a spotlight on his administration's efforts to reduce the rate of global warming through the reduction of carbon emissions. An alligator is spotted sunning itself along the Anhinga Trail at Everglades National Park as Obama took a tour. Wiping sweat from his brow as he braved the subtropical temperatures to make his mark on Earth Day, the president mingled with crowd-goers at the mid-day event before returning to Washington on Air Force One. Previewing the day trip, the White House had reiterated Obama's belief that 'no challenge poses a greater threat to future generations than climate change.' It's statement pointed out that last year was the warmest year on record and that 14 of the 15 of earth's hottest years happened this century. It said, 'Climate change is also affecting some of the most iconic places in our country, from disappearing glaciers in Glacier National Park to dying Joshua Trees in Joshua Tree National Park.' 'Recognized worldwide as a unique and treasured landscape, the Everglades is a perfect example of the threat we face from climate change, including rising sea levels that result in shoreline erosion and increased flooding.' Obama said it is this generation's 'turn to ensure that' natural monuments remain the 'birthright of all Americans for generations to come.' Wiping sweat from his brow as he braved the subtropical temperatures to make his mark on Earth Day, the president mingled with crowd-goers at the mid-day event before returning to Washington on Air Force One. The White House also asserted that climate change was already displacing animals and plants, including tropical orchids, some of which are only found in south Florida. 'We are blessed with the most beautiful God-given landscape in the world,' the president said today. 'It's an incredible bounty that's been given to us. But we've got to be good stewards for it. We have to take care of it.' Obama said it is this generation's 'turn to ensure that' natural monuments remain the 'birthright of all Americans for generations to come.' 'So many people here are active in your communities, doing what's needed. The young people who are here, the next generation, they're way ahead of us in understanding how important this is. Let's make sure we

don't disappoint them. Let's stand up and do what's right before it's too late,' he concluded. Obama was accompanied by Bill Nye 'the Science Guy' on the day trip. He also gave Florida Rep. Debbie Wasserman Schultz, who he's seen here hugging, and Interior Secretary Sally Jewell, second from left, a ride ."

Autoregressive	1. President Obama highlights shrinking Florida Everglades as proof that climate change is real and a threat to national treasures and local economies . 2. Obama's trip to Florida on Air Force One used 9,000 gallons of fuel, which he acknowledges, but the White House defends it as part of the effort to reduce emissions . 3. Obama urges Americans to take environmentally-conscious actions to save the Everglades and other natural gems.
SD	1. President Obama highlights Florida Everglades as proof of climate change impact. 2. Obama criticizes climate change as a current issue, not a future threat. 3. White House defends gas-guzzling trip to Florida for Earth Day, linking it to efforts to reduce pollutants.
SD (Lossy)	1. President Obama highlights shrinking Florida Everglades as proof climate change is real and threatening national treasures and local economies. 2. Obama's trip to Florida, which required 9,000 gallons of fuel for Air Force One, drew criticism over its environmental impact. 3. Obama emphasizes that climate change is a current issue with serious implications, not a problem for future generations. 4. Obama's administration is committed to leading the world in combating climate change through reducing carbon emissions. 5. Obama's Earth Day visit to the Everglades was meant to inspire environmentally-conscious decisions to protect natural treasures.
SpecCascade	1. President Obama highlights climate change as a threat to the Everglades and national treasures. 2. The president says climate change is not a problem for another generation. 3. Obama's trip to Florida was criticized for being gas-guzzling, but the White House said it would not undermine his Earth Day message. 4. The White House said that climate change is affecting iconic places like the Everglades and Glacier National Park. 5. Obama emphasizes the need for action to protect natural monuments as a birthright for future generations.
Static Ensemble	1) President Barack Obama visits Florida Everglades as part of Earth Day message. 2) Obama and Bill Nye 'the Science Guy' on trip. 3) Obama addresses climate change as a threat to national treasures and economies. 4) Obama mentions the impact of climate change on the Everglades and surrounding towns. 5) Obama emphasizes the need to take action on climate change.
DIVERSED	1. President Obama highlights Florida Everglades as proof of climate change's real and threatening impact. 2. Obama commits U.S. to lead in combatting climate change and reducing carbon emissions. 3. Obama's trip to Florida, accompanied by Bill Nye 'the Science Guy,' was seen as a challenge to his Earth Day message.

News: President Barack Obama held up the shrinking Florida Everglades today as proof positive that climate change is real and is threatening America's 'national treasures,' as well as the economies of their surrounding towns, which rely heavily on tourist dollars. 'Climate change can no longer be denied. It can't be edited out,' he said this afternoon in a speech at Everglades National Park commemorating Earth Day. 'It can't be omitted from the conversation. And action can no longer be delayed.' 'That's why I've committed the United States to lead the world in combatting this threat,' he said, throwing a spotlight on his administration's efforts to reduce the rate of global warming through the reduction of carbon emissions. The day trip to South Florida, on which Obama was accompanied by Bill Nye 'the Science Guy,' also highlighted the massive amount of fuel it takes to power Air Force One, however, 9,000 gallons, round trip. Scroll down for video. President Barack Obama takes a walking tour of the Anhinga Trail at Everglades National Park. Obama visited the subtropical swamps of the park today as part of a push to get Americans thinking and talking about the damage climate change is causing close to home. Obama held up the shrinking Florida Everglades today as proof positive that climate change is real and is threatening America's 'national treasures,' as well as the economies of their surrounding towns, which rely heavily on tourist dollars. 'Climate change can no longer be denied. It can't be edited out,' he said this afternoon in a speech at Everglades National Park commemorating Earth Day. 'It can't be omitted from the conversation. And action can no longer be delayed' Vehicles in the president's motorcade are seen in the Everglades National Park where Obama stopped for a visit on the Anhinga trail in Homestead. Asked earlier this week if Obama's gas-guzzling trip to Florida would undermine his Earth Day message, the White House said it would not and reaffirmed Obama's efforts to reduce pollutants. On of the helicopters escorting Marine One with President Barack Obama onboard, prepares to land at Miami's International Airport. I love the smell of jet fuel, Nye told reporters as he waited to board the president's 747 this morning for the flight to Florida, during which the White House said he would film a video with Obama. Asked earlier this week if Obama's gas-guzzling trip to Florida would undermine his Earth Day message, the White House said it would not and reaffirmed Obama's efforts to reduce pollutants. Obama took a half-hour walking tour of the the Anhinga Trail at the 1.5-million-acre national park today before making a plea for Americans to save the Everglades and other natural gems by making environmentally-conscious decisions. 'I can't think of a better way to spend Earth Day than in one of our nation's greatest natural treasures, the Everglades,' he told his audience, calling the swamp, which he acknowledged is not technically a swamp, 'magical.' 'Climate change is threatening this treasure and the communities that depend on it,' Obama warned. 'If we don't act, there may not be an Everglades as we know it.' The president said the fallout from rising temperatures is 'not a problem for another generation.' At least, 'not anymore,' he said. 'This is a problem now. It has serious implications for the way we live right now.' Obama said the rising sea level in South Florida and inward flow of salt water is evidence that climate change is already having a negative effect on communities. 'This is not some impossible problem that we cannot solve. We can solve it if we've got some political will,' he said. 'And we can solve it in a way that creates jobs. We can solve it in a way that doesn't disrupt our economy but enhances our economy. And it's a bipartisan issue.' 'That's why I've committed the United States to lead the world in combatting this threat,' he said, throwing a spotlight on his administration's efforts to reduce the rate of global warming through the reduction of carbon emissions. An alligator is spotted sunning itself along the Anhinga Trail at Everglades National Park as Obama took a tour. Wiping sweat from his brow as he braved the subtropical temperatures to make his mark on Earth Day, the president mingled with crowd-goers at the mid-day event before returning to Washington on Air Force One. Previewing the day trip, the White House had reiterated Obama's belief that 'no challenge poses a greater threat to future generations than climate change.' It's statement pointed out that last year was the warmest year on record and that 14 of the 15 of earth's hottest years happened this century. It said, 'Climate change is also affecting some of the most iconic places in our country, from disappearing glaciers in Glacier National Park to dying Joshua Trees in Joshua Tree National Park.' 'Recognized worldwide as a unique and treasured landscape, the Everglades is a perfect example of the threat we face from climate change, including rising sea levels that result in shoreline erosion and increased flooding.' Obama said it is this generation's 'turn to ensure that' natural monuments remain the 'birthright of all Americans for generations to come.' Wiping sweat from his brow as he braved the subtropical temperatures to make his mark on Earth Day, the president mingled with crowd-goers at the mid-day event before returning to Washington on Air Force One. The White House also asserted that climate change was already displacing animals and plants, including tropical orchids, some of which are only found in south Florida. 'We are blessed with the most beautiful God-given landscape in the world,' the president said today. 'It's an incredible bounty that's been given to us. But we've got to be good stewards for it. We have to take care of it.' Obama said it is this generation's 'turn to ensure that' natural monuments remain the 'birthright of all Americans for generations to come.' 'So many people here are active in your communities, doing what's needed. The young people who are here, the next generation, they're way ahead of us in understanding how important this is.' 'Let's make sure we

don't disappoint them. Let's stand up and do what's right before it's too late,' he concluded. Obama was accompanied by Bill Nye 'the Science Guy' on the day trip. He also gave Florida Rep. Debbie Wasserman Schultz, who he's seen here hugging, and Interior Secretary Sally Jewell, second from left, a ride."

Autoregressive	President Obama held up the shrinking Florida Everglades today as proof positive that climate change is real and is threatening America's 'national treasures'. Obama says climate change is already having a negative effect on communities. The president warned that if action isn't taken, there may not be an Everglades as we know it. He was accompanied by Bill Nye the Science Guy on the day trip.
SD	9,000 gallons of fuel used for Air Force One round trip . Obama held up the shrinking Florida Everglades as proof positive that climate change is real . Obama said it is this generation's 'turn to ensure that' natural monuments remain the 'birthright of all Americans for generations to come.'
SD (Lossy)	9,000 gallons of fuel used for Air Force One round trip to Florida . Obama held up the shrinking Florida Everglades as proof positive that climate change is real . President said the rising sea level in South Florida and inward flow of salt water is evidence that climate change is already having a negative effect on communities . Obama said it is this generation's 'turn to ensure that' natural monuments remain the 'birthright of all Americans for generations to come.'
SpecCascade	9,000 gallons of fuel used for Air Force One round trip . Obama held up the shrinking Florida Everglades as proof positive that climate change is real . Obama said it is this generation's 'turn to ensure that' natural monuments remain the 'birthright of all Americans for generations to come.'
Static Ensemble	1. President Obama held up the shrinking Florida Everglades as proof of climate change. 2. Obama's trip to Florida included a walking tour of the Everglades and a speech on Earth Day. 3. The trip highlighted the administration's efforts to reduce carbon emissions. 4. Obama's gas-guzzling trip to Florida was criticized, but the White House defended it. 5. Obama emphasized the urgency of addressing climate change and its impact on communities."
DIVERSED	1. Obama uses Everglades as proof of climate change's impact. 2. Trip highlights shrinking Everglades and economic impact on surrounding towns. 3. Obama's Air Force One trip uses 9,000 gallons of fuel, raising questions about hypocrisy. 4. Bill Nye 'the Science Guy' accompanies Obama. 5. Obama emphasizes the urgency of addressing climate change and its impact on future generations."

C.6 XSum Case Study

Case Study on the XSum dataset using Llama model.

Gao Yu, 71, was found guilty last April and challenged her conviction at a closed hearing in Beijing on Thursday. The guilty verdict was upheld, but the jail term was cut from seven years to five and the medical grounds permitted. Foreign governments and rights groups have condemned the guilty verdict, calling it politically motivated. Sources have told the BBC that Ms Gao has already left prison. The well-known investigative journalist was found guilty of sending an internal Communist Party document to Mingjing News, a Chinese-language website in the United States. The document warned against the dangers of a free press and an independent civil society. At the time of the original conviction the court had said that Ms Gao had illegally provided state secrets to foreigners. Both Ms Gao and Mingjing News had denied she was the source of the leak. However, on Thursday, Ms Gao confessed to the crime and expressed regret for her actions, and the appeals court reduced her sentence to five years and decided Ms Gao could temporarily serve the sentence out of prison as she was severely ill, Chinese state media said. Experts say Chinese courts often grant reduced sentences after convicts plead guilty. However, Ms Gao's lawyers told AP they continued to argue that she was innocent. Many believe that Gao Yu is unpopular with the Chinese government because of her unrelenting reports focusing on China's elite-level politics, the BBC's Celia Hatton in Beijing reports. Many people had pressed for a lighter sentence in consideration of Ms Gao's advanced age and her chronic heart problems. Her son, Zhao Meng, had told the BBC he was concerned his mother would not survive another lengthy period in prison, her third since 1989. Prominent human rights activist Hu Jia said (in Chinese): 'I'm relieved Gao Yu can return to her own home, after a year and a half in custody... but the risk of being returned to jail at any time still hangs over Gao Yu's head, like the sword of Damocles.'

Autoregressive	Chinese journalist Gao Yu has had her seven-year prison sentence cut to five years and allowed to serve at home. Her sentence was reduced after she confessed to the crime and expressed regret.
SD	A Chinese journalist's conviction for leaking state secrets is upheld, but her jail term is reduced to five years. She is released from prison due to health reasons. She confessed to the crime. She is still at risk of being returned to jail. She is a prominent investigative journalist. She was found guilty of sending an internal Communist Party document to a Chinese-language website in the United States.
SD (Lossy)	A Chinese journalist's conviction for leaking state secrets is upheld, but her jail term is reduced to five years. She is released from prison due to health reasons. She confessed to the crime. She is still at risk of being returned to jail. She is a prominent investigative journalist. She was found guilty of sending an internal Communist Party document to a Chinese-language website in the United States.
SpecCascade	A Chinese journalist's conviction for leaking state secrets is upheld, but her jail term is reduced to five years. She is released from prison due to health reasons. She confessed to the crime. She is still at risk of being returned to jail. She is a prominent investigative journalist. She was found guilty of sending an internal Communist Party document to a Chinese-language website in the United States. She was found guilty of illegally providing state secrets to foreigners.
Static Ensemble	71-year-old Chinese investigative journalist Gao Yu's conviction is upheld, but her sentence is reduced from 7 years to 5 years. Gao Yu, a 71-year-old Chinese journalist, is released from prison after her conviction is upheld. Gao Yu's conviction for leaking state secrets is upheld but her sentence is reduced to 5 years. Chinese authorities reduce Gao Yu's 7-year sentence to 5 years, allowing her to leave prison early due to her age and health. Gao Yu's sentence is reduced to 5 years, allowing her to return home after serving a year and a half in custody.
DIVERSED	71-year-old Chinese journalist Gao Yu has been released from prison after her conviction was upheld, but her sentence was reduced to five years. She was found guilty of leaking state secrets. Gao Yu has been released from prison after her conviction was upheld, but her sentence was reduced to five years. She was found guilty of leaking state secrets. Gao Yu has been released from prison after her conviction was upheld, but her sentence was reduced to five years.

Case Study on the XSum dataset using Gemma model.

Gao Yu, 71, was found guilty last April and challenged her conviction at a closed hearing in Beijing on Thursday. The guilty verdict was upheld, but the jail term was cut from seven years to five and the medical grounds permitted. Foreign governments and rights groups have condemned the guilty verdict, calling it politically motivated. Sources have told the BBC that Ms Gao has already left prison. The well-known investigative journalist was found guilty of sending an internal Communist Party document to Mingjing News, a Chinese-language website in the United States. The document warned against the dangers of a free press and an independent civil society. At the time of the original conviction the court had said that Ms Gao had illegally provided state secrets to foreigners. Both Ms Gao and Mingjing News had denied she was the source of the leak. However, on Thursday, Ms Gao confessed to the crime and expressed regret for her actions, and the appeals court reduced her sentence to five years and decided Ms Gao could temporarily serve the sentence out of prison as she was severely ill; Chinese state media said. Experts say Chinese courts often grant reduced sentences after convicts plead guilty. However, Ms Gao's lawyers told AP they continued to argue that she was innocent. Many believe that Gao Yu is unpopular with the Chinese government because of her unrelenting reports focusing on China's elite-level politics, the BBC's Celia Hatton in Beijing reports. Many people had pressed for a lighter sentence in consideration of Ms Gao's advanced age and her chronic heart problems. Her son, Zhao Meng, had told the BBC he was concerned his mother would not survive another lengthy period in prison, her third since 1989. Prominent human rights activist Hu Jia said (in Chinese): 'I'm relieved Gao Yu can return to her own home, after a year and a half in custody... but the risk of being returned to jail at any time still hangs over Gao Yu's head, like the sword of Damocles.'

Autoregressive	A Chinese journalist, Gao Yu, had her conviction upheld, jail term reduced, and released from prison on medical grounds .
SD	A Chinese journalist Gao Yu's conviction was upheld, sentence reduced, and she has been released from prison on medical grounds .
SD (Lossy)	A Chinese journalist's conviction was upheld, her sentence reduced, and she has been released from prison on medical grounds .
SpecCascade	A Chinese journalist's conviction was upheld, but her sentence was reduced and she was released due to illness and a guilty plea.
Static Ensemble	Chinese journalist Gao Yu's conviction was upheld, sentence reduced, and she has been released from prison due to health concerns .
DIVERSED	Gao Yu, a journalist, was convicted of leaking state secrets, but her sentence was reduced and she has been released from prison.

Case Study on the XSum dataset using Qwen model.

Gao Yu, 71, was found guilty last April and challenged her conviction at a closed hearing in Beijing on Thursday. The guilty verdict was upheld, but the jail term was cut from seven years to five and the medical grounds permitted. Foreign governments and rights groups have condemned the guilty verdict, calling it politically motivated. Sources have told the BBC that Ms Gao has already left prison. The well-known investigative journalist was found guilty of sending an internal Communist Party document to Mingjing News, a Chinese-language website in the United States. The document warned against the dangers of a free press and an independent civil society. At the time of the original conviction the court had said that Ms Gao had illegally provided state secrets to foreigners. Both Ms Gao and Mingjing News had denied she was the source of the leak. However, on Thursday, Ms Gao confessed to the crime and expressed regret for her actions, and the appeals court reduced her sentence to five years and decided Ms Gao could temporarily serve the sentence out of prison as she was severely ill, Chinese state media said. Experts say Chinese courts often grant reduced sentences after convicts plead guilty. However, Ms Gao's lawyers told AP they continued to argue that she was innocent. Many believe that Gao Yu is unpopular with the Chinese government because of her unrelenting reports focusing on China's elite-level politics, the BBC's Celia Hatton in Beijing reports. Many people had pressed for a lighter sentence in consideration of Ms Gao's advanced age and her chronic heart problems. Her son, Zhao Meng, had told the BBC he was concerned his mother would not survive another lengthy period in prison, her third since 1989. Prominent human rights activist Hu Jia said (in Chinese): 'I'm relieved Gao Yu can return to her own home, after a year and a half in custody... but the risk of being returned to jail at any time still hangs over Gao Yu's head, like the sword of Damocles.'

Autoregressive	71-year-old Chinese journalist Gao Yu was sentenced to five years in prison for leaking a Party document but was allowed to serve it outside prison due to illness.
SD	71-year-old Chinese journalist Gao Yu was convicted of leaking state secrets, received a five-year sentence, and is temporarily released due to illness.
SD (Lossy)	71-year-old Chinese journalist Gao Yu was convicted of leaking state secrets, received a reduced sentence, and is now out of prison due to illness.
SpecCascade	71-year-old Chinese journalist Gao Yu was convicted of leaking state secrets, received a reduced sentence, and is now out of prison due to illness.
Static Ensemble	71-year-old Chinese journalist Gao Yu was found guilty but received a reduced sentence, and the court allowed her to temporarily serve her sentence out of prison due to her serious illness.
DIVERSED	71-year-old Chinese journalist Gao Yu was convicted of leaking state secrets, received a five-year sentence, and is now allowed to leave prison temporarily due to illness, despite her lawyers' claims of innocence.

C.7 MBPP Case Study

Case Study on the MBPP dataset using Llama model.

Write a function to find the median of three numbers.

Autoregressive	<pre>def median_numbers(a,b,c): numbers = sorted([a,b,c]) if len(numbers) return (numbers[len(numbers)//2 - 1] + numbers[len(numbers)//2]) / 2 else: return numbers[len(numbers)//2]</pre>
----------------	--

SD	<pre>def median_numbers(a,b,c): numbers = sorted([a,b,c]) if len(numbers) return (numbers[len(numbers)//2 - 1] + numbers[len(numbers)//2]) / 2 else: return numbers[len(numbers)//2]</pre>
----	--

SD (Lossy)	<pre>def median_numbers(a,b,c): return sorted([a,b,c])[1]</pre>
------------	---

SpecCascade	<pre>def median_numbers(a,b,c): numbers = sorted([a,b,c]) if len(numbers) median = (numbers[len(numbers)//2 - 1] + numbers[len(numbers)//2]) / 2 else: median = numbers[len(numbers)//2] return median</pre>
-------------	--

Static Ensemble	<pre>def median_numbers(a,b,c): return sorted([a,b,c])[1]</pre>
-----------------	---

DIVERSED	<pre>def median_numbers(a,b,c): return sorted([a,b,c])[1]</pre>
----------	---

Case Study on the MBPP dataset using Qwen model.

Write a function to find the median of three numbers.

Autoregressive	<code>def median_numbers(a,b,c): return sorted([a,b,c])[1]</code>
SD	<code>def median_numbers(a,b,c): return sorted([a,b,c])[1]</code>
SD (Lossy)	<code>def median_numbers(a,b,c): return sorted([a,b,c])[1]</code>
SpecCascade	<code>def median_numbers(a,b,c): return sorted([a,b,c])[1]</code>
Static Ensemble	<code>def median_numbers(a,b,c): return sorted([a,b,c])[1]</code>
DIVERSED	<code>def median_numbers(a,b,c): return sorted([a,b,c])[1]</code>

Case Study on the MBPP dataset using Gemma model.

Write a function to find the median of three numbers.

Autoregressive	<pre>def median_numbers(a,b,c): if (a <= b and b <= c) or (c <= b and b <= a): return b elif (b <= a and a <= c) or (c <= a and a <= b): return a else: return c</pre>
SD	<pre>def median_numbers(a,b,c): numbers = sorted([a,b,c]) if len(numbers) return (numbers[len(numbers)//2 - 1] + numbers[len(numbers)//2]) / 2 else: return numbers[len(numbers)//2]</pre>
SD (Lossy)	<pre>def median_numbers(a,b,c): if (a <= b and b <= c) or (c <= b and b <= a): return b elif (b <= a and a <= c) or (c <= a and a <= b): return a else: return c</pre>
SpecCascade	<pre>def median_numbers(a,b,c): if (a <= b and b <= c) or (c <= b and b <= a): return b elif (b <= a and a <= c) or (c <= a and a <= b): return a else: return c</pre>
Static Ensemble	<pre>def median_numbers(a,b,c): if (a <= b and b <= c) or (c <= b and b <= a): return b elif (b <= a and a <= c) or (c <= a and a <= b): return a else: return c</pre>
DIVERSED	<pre>def median_numbers(a,b,c): if (a <= b and b <= c) or (c <= b and b <= a): return b elif (b <= a and a <= c) or (c <= a and a <= b): return a else: return c</pre>