# DIVERSED: Relaxed Speculative Decoding via Dynamic Ensemble Verification

**Ziyi Wang**[*,1], **Siva Rajesh Kasa**[*,2], **Ankith M S**[2], **Santhosh Kumar Kasa**[2], **Jiaru Zou**[3], **Sumit Negi**[2], **Ruqi Zhang**[1], **Nan Jiang**[4], **Qifan Song**[1]
[1]Purdue University, [2]Amazon Inc., [3]University of Illinois Urbana-Champaign, [4]University of Texas at El Paso

## Abstract

Speculative decoding is an effective technique for accelerating large language model inference by drafting multiple tokens in parallel. In practice, its speedup is often bottlenecked by a rigid verification step that strictly enforces the accepted token distribution to exactly match the target model. This constraint leads to the rejection of many plausible tokens, lowering the acceptance rate and limiting overall time speedup. To overcome this limitation, we propose DynamIc VERification RElaxed SpEculative Decoding (DIVERSED), a relaxed verification framework that improves time efficiency while preserving generation quality. DIVERSED: learns an ensemble-based verifier that blends the draft and target model distributions with a task-dependent and context-dependent weight. We provide theoretical justification for our approach and demonstrate empirically that DIVERSED achieves substantially higher inference efficiency compared to standard speculative decoding methods. Code is available at: `https://github.com/comeusr/diversed`.

## 1 INTRODUCTION

Recent advances in large language models have demonstrated strong capabilities on demanding tasks, including mathematical reasoning (Kojima et al., 2022; Wei et al., 2022; Zou et al., 2025b), complex programming (Rozière et al., 2023; Yan et al., 2025; Jiang et al., 2024), and tool use (Schick et al., 2023; Qin et al., 2024; Zou et al., 2025a). Yet, these gains are achieved with billion-scale models, which substantially increase inference cost and make efficiency a central bottleneck. Existing efforts have improved runtime via faster attention (Dao et al., 2022), model compression (Dettmers et al., 2022), and accelerating decoding (Leviathan et al., 2023; Narasimhan et al., 2025).

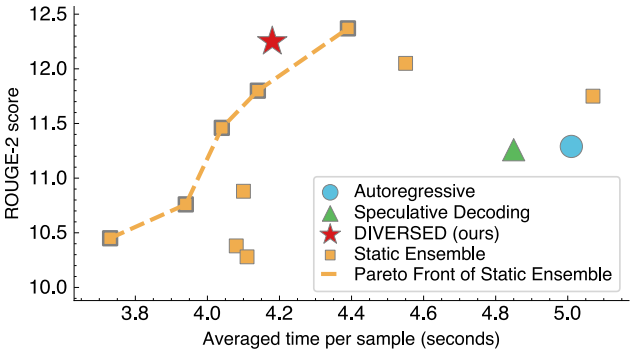


Figure 1: Our method improves upon the Pareto frontier achieved by static ensemble, attaining better trade-offs between inference time and generation quality. Results are reported on the CNNDM dataset using the target/draft model pair: Llama-3.1-8B-Instruct and Llama-3.2-1B-Instruct, with temperature being 0. More results are in Appendix Figure 10.

Among decoding strategies, speculative decoding has emerged as a particularly effective approach (Leviathan et al., 2023; Chen et al., 2023). It employs a small draft model to propose tokens and a large target model to verify them in parallel, reducing latency while maintaining target-level quality. Subsequent research has progressed along two main directions: (1) lossless methods that preserve exact agreement with the target model, and (2) lossy methods that accept more draft tokens at the cost of potential quality degradation. Lossless work includes self-speculative architectures that restructure or internalize the draft process (Cai

[*]Equal contribution.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

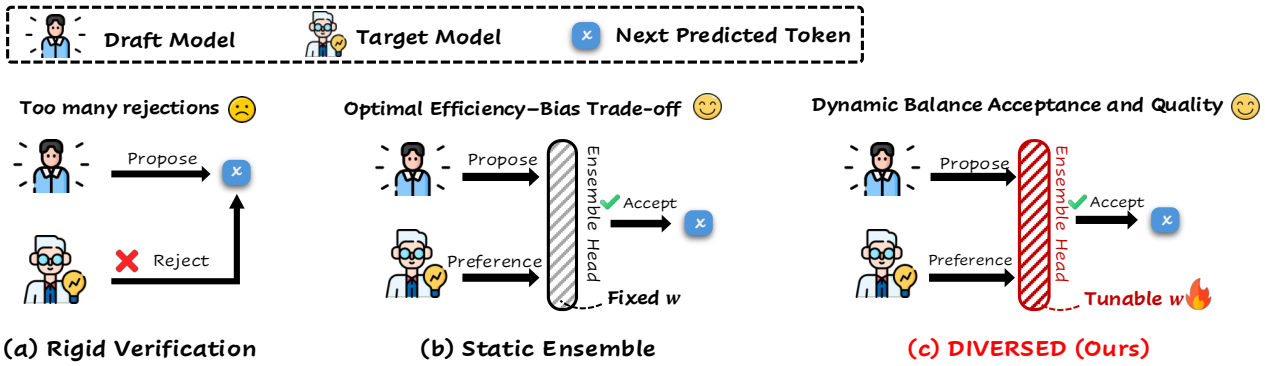


Figure 2: Compared with classic speculative decoding (shown in (a) rigid verification), our DIVERSED (c) achieves a higher acceptance rate, comparable accuracy, and lower wall-clock time. Compared with static ensemble (b), DIVERSED (c) attains higher accuracy via a tunable ensemble weight that adapts to the task and the context.

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., 2024). On the other hand, lossy approaches intentionally relax token-level agreement with the target, trading performance degradation for higher acceptance and throughput (Bachmann et al., 2025). Recent theory further formalizes this acceptance rate and generation quality tradeoff (Yin et al., 2024).

This work follows the lossy speculative decoding line, aiming to increase acceptance while incurring minimal degradation in generation quality. This is feasible because many tokens rejected by the target verifier can still lead to correct solutions (Yan et al., 2025; Bachmann et al., 2025); rejecting them outright can limit overall time speedup. An intuitive idea is to relax the verification rule by interpolating the draft and target distributions using a static weight, noted as static ensemble. It directly traces the Pareto frontier characterized by Yin et al. (2024). Figure 3 illustrates both a success case and a failure case. While a static ensemble can improve acceptance without harming quality in some instances, it can also degrade generation quality and yield incorrect solutions. This suggests that the ensemble rule should be context-dependent, rather than fixed across all generated contexts.

To address this limitation, we propose DynamIc VErification RElaxed SpEculative Decoding (DIVERSED), a dynamic ensemble verifier that mixes the target and draft distributions with context- and token-dependent weights. These weights are learned by optimizing a regularized objective that maximizes expected task reward on the dataset while encouraging a high acceptance rate. As shown in Figure 1, DIVERSED achieves higher acceptance at comparable quality, surpassing the Pareto frontier of static ensemble. A visual comparison of DIVERSED against existing baselines is shown in Figure 2.

Theoretically, we show that (1) the *static ensemble* verifier lies on the acceptance-quality Pareto front proposed by Yin et al. (2024). Importantly, this implies that practitioners need not 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. (2) We also derive an exact, step-dependent expression for the expected accepted length. This characterization captures position effects overlooked in prior analyses and eliminates both the *i.i.d.* assumption adopted in earlier work. This extension is technically non-trivial.

Experimentally, we show (1) DIVERSED consistently increases acceptance rates, while preserving task accuracy across all model pairs and task datasets. (2) The wall-clock speedup closely correlated to acceptance rate: increases in acceptance rate reliably translate into end-to-end latency reductions. (3) The efficiency–quality trade-off exhibits a clear Pareto front, on which DIVERSED consistently achieves a strong balance between speed and accuracy.

## 2 PRELIMINARY

**Speculative decoding.** In autoregressive language models, a target model generates tokens $(x_1, \ldots, x_T)$ from a prompt $x_0$, by sampling tokens one at a time from its distribution $p(\cdot|x_{0:t-1})$, Speculative decoding accelerates this process by introducing a smaller, faster draft model $q(\cdot|x_{0:t-1})$ that proposes a block of $N$ draft tokens $\tilde{x}_{t:t+N-1}$. For convenience, we write $p_t(\cdot) = p(\cdot|x_{0:t-1})$ and $q_t(\cdot) = q(\cdot|x_{0:t-1})$, when the prefix $x_{0:t-1}$ is clear from the context. The target model then verifies in parallel by computing $p(\tilde{x}_{t:t+N-1}|x_{0:t-1})$

and accepts $\tilde{x}_t$ with probability

$$b_t(\tilde{x}_t) \;=\; \min\{1,\; p_t(\tilde{x}_t)/q_t(\tilde{x}_t)\}\,.$$

If accepted, we set $x_t$ as $\tilde{x}_t$. Otherwise, the target samples a replacement token $x_t$ from the fallback distribution

$$P_t(x) = \frac{[p_t(x) - q_t(x)]_+}{\sum_z [p_t(z) - q_t(z)]_+},$$

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 $T_N$ denote the number of accepted tokens in a block of $N$ draft tokens. Prior work (Leviathan et al., 2023) shows that $\mathbb{E}[T_N] = (1 - \beta^{N+1})/(1 - \beta)$, where $\beta$ is the expected accepted probability and equals to $1 - \mathrm{TV}(p, q)$. These guarantees, however, rely on the assumption that token verification is independent and identical (*i.i.d.*), which have not considered the effect of the draft token positions in the draft sequence.

**Lossy speculative decoding and its Pareto front.** A *lossy* variant of speculative decoding replaces the standard rule with a relaxed accept rule $b_t(\tilde{x}_t)$. Upon rejection, it samples from a fallback $P_t(\cdot)$, inducing a relaxed token distribution $\mathbb{P}(\cdot|b_t, P_t)$. The detailed procedure is given in Algorithm 1. To increase acceptance relative to lossless speculative decoding, the accept rule should be more permissive than the standard rule. The design objective of lossy speculative decoding is a multi-objective problem that simultaneously minimizes (1) the rejection probability $\mathbb{P}_{b_t}(\text{reject})$ and (2) the total variation bias $\mathrm{TV}(\mathbb{P}(\cdot; b_t, P_t),\, p)$. We formalize the trade-off between acceptance and distribution bias via Pareto optimality.

**Definition 1** (Pareto-front solutions for lossy speculative decoding)**.** *We say that an accept–fallback pair* $(b_t, P_t)$ dominates $(b'_t, P'_t)$ *if* $\mathbb{P}_{b_t}(\text{reject}) < \mathbb{P}_{b'_t}(\text{reject})$ *and* $\mathrm{TV}(\mathbb{P}(\cdot|b_t, P_t),\, p) \;<\; \mathrm{TV}\left(\mathbb{P}(\cdot; b'_t, P'_t),\, p\right)$*. A pair* $(b_t, P_t)$ *is* Pareto-optimal *(non-dominated) if no other pair dominates it. The set of all Pareto-optimal pairs forms the* Pareto front*.*

**Theorem 1** (Yin et al. (2024), Theorem 5; informal)**.** *An accept–fallback pair* $(b_t, P_t)$ *is Pareto-optimal if and only if it satisfies*

$$\mathbb{P}_{b_t}(\text{reject}) + \mathrm{TV}(\mathbb{P}(\cdot; b_t, P_t),\, p) \;=\; \mathrm{TV}(q,\, p). \qquad (1)$$

Theorem 1 shows that reducing the rejection probability—and thus improving time efficiency—necessarily comes at the cost of increased distribution bias. Here, *distribution bias* is the discrepancy between the token distribution induced by relaxed speculative decoding and the target-model distribution, measured (following prior work (Yin et al., 2024)) by the total variation distance.

# 3 METHODOLOGY

## 3.1 Static Ensemble

**Motivating example.** Figure 4 (a) studies the relation between inference time and acceptance rate. The results are obtained by generating 128 tokens on the CNNDM dataset, using three target/draft pairs from the Llama, Qwen, and Gemma families. Across all pairs, we observe that wall-clock time decreases monotonically as acceptance rate increases. This highlights acceptance as the primary driver of speedup.

Motivated by this observation, a straightforward idea 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 weight $0 < w \leq 1$, define the ensemble distribution $\nu_t$ as:

$$\nu_t(\cdot) = w p_t(\cdot) + (1 - w) q_t(\cdot). \qquad (2)$$

During verification, 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:

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

Following the analysis of 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 $\mathbb{P}(x_t = x \mid x_{0:t-1}) = \nu_t(x)$.

**Proposition 1.** *The static ensemble achieves the Pareto front, i.e.,* $(\nu_t, P_{\nu_t})$ *satisfies:*

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

*Proof is deferred to Appendix B.1.*

Together with Theorem 1, proposition 1 demonstrates that the static ensemble exactly characterizes the Pareto-optimal tradeoff between acceptance probability and distributional bias. By varying the ensemble weight $w$, $(\nu_t, P_{\nu_t})$ traverses the entire Pareto front.

We then quantify how the static ensemble increases the probability of accepting a single draft token.

**Lemma 1.** *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*

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

*The proof is deferred to Appendix B.2.*

Lemma 1 shows that static ensemble increases the acceptance probability from $1 - \mathrm{TV}(q, p)$ to $1 - w\ \mathrm{TV}(q, p)$. When $w = 1$, it recovers the lossless speculative decoding; when $w < 1$, it boosts acceptance.

[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 day, and
[and] there ...✓ [So], each bird...✓
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 + ... + \$18 = \$143 on the items she has receipts for. She has \$16 left from her budget, so
[\$153] on the ... she spent \$49. ×
she spent ... \$184 in total. Therefore, she spent \$184 − \$143 = \$41 on the shoes.
[on] the shoes ...×

Figure 3: Examples of speculative-decoding verification with accepted mismatches, illustrating not all draft–target mismatches are equally harmful. **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.

### 3.2 Dynamic Verification Relaxed Speculative Decoding

**Motivating example.** Figure 3 illustrates speculative-decoding verification under accepted mismatches. The example highlights that relaxing verification at different positions can have markedly different effects on generation quality: not all draft–target mismatches are equally harmful. Tokens marked in green can be relaxed without degrading the final output, whereas relaxing the red tokens leads to an incorrect answer. This observation suggests that verification relaxation should be context-dependent. To increase acceptance while preserving correctness, the verifier should lean toward the target distribution for quality-critical tokens (e.g., tokens that can derail a mathematical solution), while allowing more aggressive relaxation—and thus higher draft acceptance—for less influential tokens.

We introduce a weighting scheme, where the weight $w_t$ is conditioned on the context $x_{0:t-1}$ at each time step $t$. To represent the 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 $h_t^p$, which together summarize $x_{0:t-1}$. The ensemble weight is parameterized by a feed-forward neural network:

$$w_t = f_\theta(h_t^q, h_t^p). \tag{3}$$

The resulting dynamic ensemble distribution becomes

$$\nu_t^\theta(x) = w_t p_t(x) + (1 - w_t) q_t(x). \tag{4}$$

During verification, a draft token $\tilde{x}_t$ is accepted with this new probability $b_{\nu_t}(x) = \min\{1, \nu_t^\theta(x)/q_t(x)\}$. Similar to Section 3.1, the resulting token distribution matches the intended ensemble distribution $\mathbb{P}(x_t = x \mid x_{0:t-1}) = \nu_t^\theta(x)$.

**Training.** During training, the draft and target models process the same prompt, and the ensemble distribution $\nu_t^\theta$ is computed as in Equation (4), from which the next token $x_t$ is sampled. Token-level supervision is typically unavailable (Bachmann et al., 2025), so we train $\nu^\theta$ with reinforcement learning using a sequence-level reward $R(x_{1:T})$. For example, in mathematical reasoning, $R = 1$ if the final answer is correct and $R = 0$ otherwise. Let $D$ be the dataset of prompts $x_0$. We optimize a regularized objective that balances task performance and acceptance:

$$\mathbb{E}_{x_0 \sim D,\, x_{1:T} \sim \nu^\theta(x_0)} \big[ R(x_{1:T}) + \lambda \sum_{t=0}^{T} \big(1 - \mathrm{TV}(q_t, \nu_t^\theta)\big) \big], \tag{5}$$

where the regularizer $\sum_{t=0}^{T} \big(1 - \mathrm{TV}(q_t, \nu_t^\theta)\big)$ promotes higher acceptance and $\lambda > 0$ controls its strength. This regularization also discourages degenerate solutions that collapse to using only the target model ($w = 0$) or only the draft model ($w = 1$). We optimize objective (5) using the recent REINFORCE++ algorithm (Hu et al., 2025), which stabilizes training through clipping and global batch averaging.

**Inference.** At inference, the draft model autoregressively propose $N$ 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 evaluates these tokens in parallel to produce $h_{t+1:t+N}^p$ and $p_{t:t+N-1}$. The ensemble head then computes weights $w_{t:t+N-1}$ via Equation (3) and the corresponding dynamic ensemble distributions $\nu_{t:t+N-1}^\theta(\cdot)$ via Equation (4) in parallel. Thus, the relaxed accept rule and fallback distribution are computed as follows:

$$\begin{aligned} b_{\nu_t^\theta}(\tilde{x}_t) &= \min\left\{1, \nu_t^\theta(\tilde{x}_t)/q_t(\tilde{x}_t)\right\}, \\ P_{\nu_t^\theta}(x) &= \frac{[\nu_t^\theta(x) - q_t(x)]_+}{\sum_z [\nu_t^\theta(z) - q_t(z)]_+}. \end{aligned} \tag{6}$$

The inference pipeline is in Appendix Algorithm 2.

**Algorithm 1** General Framework for Lossy Speculative Decoding.

**Inputs:** Maximum generation length $T$; draft length $N$; prompt $x_0$; draft model distributions $\{q_t\}$; target model distributions $\{p_t\}$; acceptance rule $b_t$; fallback distribution $\{P_t\}$.

Initialize $n \leftarrow 1$;
**while** $n \leq T$ **do**
  **for** $t = n, \ldots, n+N$ **do**
    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
  **for** $t = n, \ldots, T$ **do**
    compute $p_t(\cdot \mid x_{0:n-1}, \tilde{x}_{n:t-1})$. ▷ target parallel verification (Section 3)
  **for** $t = n \ldots n+N$ **do**
    draw $u \sim \text{Uniform}(0,1)$.
    **if** $u \leq b_t(\tilde{x}_t \mid x_{1:n-1}, \tilde{x}_{n:t-1})$ **then**
      **accept** $\tilde{x}_t$; set $x_n \leftarrow \tilde{x}_t$; update $n \leftarrow n+1$.
    **else**
      **reject**; sample $x_n \sim P_t$; update $n \leftarrow n+1$; **break**. ▷ discard $\tilde{x}_{t+1:n+N-1}$ and resume
**return** sampled sequence $x_{1:T}$.

## 4 THEORETICAL ANALYSIS

This section derives the exact expected acceptance rate given a draft length $N$. As each accepted token avoids an expensive call to the target model, the key factor driving speedup in speculative decoding is the number of accepted draft tokens (Leviathan et al., 2023). However, this was only approximated under an *i.i.d.* assumption on acceptance probabilities. In practice, this assumption is violated: the acceptance probability of a draft token is highly related to the position of the token in the draft sequence (Huang et al., 2025; Mamou et al., 2024). We therefore 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 the history $x_{0:t-1}$; otherwise $A_t = 0$. The acceptance length $T_N$ is the number of consecutive draft tokens accepted before the first rejection. When all $N$ are accepted, $T_N = N$. Next, we characterize the distribution of the accepted tokens, which will be used for computing the expected accepted length.

**Lemma 2** (Distribution of Accepted Tokens)**.** *Consider a draft and target model with corresponding conditional distributions $q(\cdot|x_{0:t-1})$ and $p(\cdot|x_{0:t-1})$ with prefix $x_{0:t-1}$. Given that a token is accepted by speculative decoding, its distribution follows the min-kernel:*

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

*Proof is deferred to Appendix B.3.*

Lemma 2 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 2; the details can be found in the Appendix C.2.

Yin et al. (2024) quantified that the number of rejections is $\sum_{t=1}^{N} \mathbb{E}_{x_{1:t-1}\sim q}[\text{TV}(q(\cdot|x_{0:t-1}), p(\cdot|x_{0:t-1}))]$. However, this view ignores the position of rejection. 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 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 ($T_N = 1$ versus $T_N = N-1$), leading to different wall-clock times. Motivated by this, we derive a more precise expression for the expected acceptance length.

**Theorem 2** (Expected Acceptance Length)**.** *Consider a draft and target model with corresponding conditional distributions $q(\cdot|x_0)$, $p(\cdot|x_0)$. Given draft length $N$, the expected acceptance length $\mathbb{E}[T_N]$ is*

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

*where the expectation is taken w.r.t. the min-kernel $\tilde{m}$ in Lemma 2.*

*Proof Sketch.* We directly characterize $\mathbb{E}[T_N|\tilde{x}_{1:N}]$ by tower law. As the fact that $T_N$ is non-negative,

we know that $\mathbb{E}[T_N|\tilde{x}_{1:N}] = \sum_{t=1}^{N} \mathbb{P}(T_N > t|\tilde{x}_{1:t-1})$, and $\mathbb{P}(T_N > t|\tilde{x}_{1:t-1})$ 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 $\mathbb{P}(T_N > t|\tilde{x}_{1:t-1})$, as it can be written as $P(\text{Accept}\,\tilde{x}_t|\tilde{x}_{1:t-1}\,\text{is accepted}) \times P(\text{Accept}\,\tilde{x}_{1:t-1}|\tilde{x}_{1:t-1})$, where the min-kernel distribution $\tilde{m}$ characterizes the accepted tokens. The full proof is presented in Appendix B.4. □

Theorem 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 $\tilde{x}_t$ exponentially decrease with the time step $t$.

# 5 RELATED WORKS

**Speculative decoding (SD)** 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× to 3× 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., 2025) 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., 2025) 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 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.

# 6 EXPERIMENTS

In experiments, 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: 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

For the choice of LLM models, we consider three target / draft model pairs: (1) Llama-3.1-8B-Instruct / Llama-3.2-1B-Instruct (Dubey et al., 2024), (2) Qwen3-8B / Qwen3-0.6B (Yang et al., 2025), (3) Gemma-3-12B-It/Gemma-3-4B-It (Team et al., 2025).

In terms of benchmark datasets, we consider: (1) GSM8K for mathematical reasoning (Cobbe et al., 2021), (2) CNNDM for news summarization (See et al., 2017), (3) XSum for extreme summarization (Narayan et al., 2018), and (4) MBPP for Python programming (Austin et al., 2021).

For the choice of baselines, we include (1) lossless speculative decoding methods, including standard speculative decoding (Leviathan et al., 2023; Chen et al., 2023) and static ensemble, and (2) representative lossy speculative decoding methods, including SD (Lossy) (Leviathan et al., 2023) and SpecCascade (Narasimhan et al., 2025). Additional experimental details are provided in Appendix C.

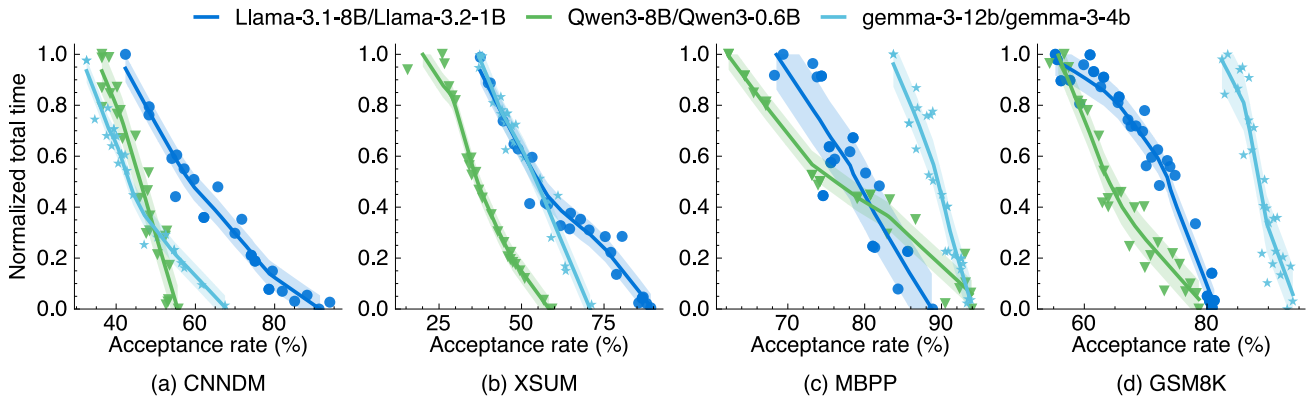


Figure 4: Normalized total time (lower is better) versus acceptance rate (higher is better) under speculative decoding with draft length $N$=5. Model pairs with higher acceptance achieve lower decoding time. Results on the CNNDM, XSUM, MBPP, and GSM8K datasets for three target/draft pairs. Additional results for draft lengths $N$=3 and $N$=7 are provided in Appendix Figures 8 and 9.

Table 1: Benchmark acceptance rate and generation quality. Columns group results by task while panels (a)–(c) correspond to different *Target*/*Draft* model pairs. All experiments use temperature 1 and draft length $N = 5$. Our DIVERSED achieves a higher acceptance rate and also maintains comparable quality. Additional results for temperatures 0 and 0.1 are reported in Appendix Tables 6 and 7.

| **Method** | **GSM8K** | | **CNNDM** | | **XSum** | | **MBPP** | |
|---|---|---|---|---|---|---|---|---|
| | Accept Rate (↑) | Quality Accuracy (↑) | Accept Rate(↑) | Quality ROUGE-2 (↑) | Accept Rate(↑) | Quality ROUGE-2 (↑) | Accept Rate(↑) | Quality pass@1 (↑) |
| 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.

### 6.2 Experimental Result and Analysis

**Benchmark inference 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 Table 1, we report the acceptance rate

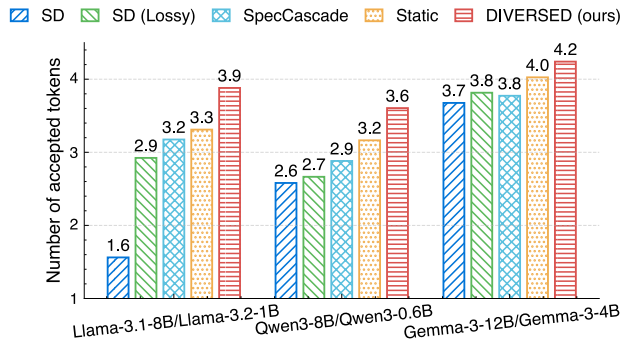


Figure 5: DIVERSED consistently accepts a larger average number of draft tokens per verification round than all baselines (draft length $N$=5, temperature 1). Appendix Figure 7 shows result for temperature 0.1.

along with the generation quality across different model pairs and various tasks. For the baselines, lossy speculative decoding, SpecCascade, and static ensemble are reported such that the generation quality matches the target model quality.

Overall, DIVERSED consistently achieves a higher 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. 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.** Figure 5 summarizes speedup by reporting the average number of draft tokens accepted per verification round for each *Target*/*Draft* pair, averaged over all datasets and generations.

DIVERSED achieves the highest accepted-token count across all pairs, reflecting its adaptive, token- and context-dependent blending of the draft and target distributions: it expands the acceptance region for high-confidence draft tokens while controlling distributional deviation from the target.

**Benchmark inference time and generation quality.** Figure 1 plots summarization quality (ROUGE-2) against end-to-end inference time. Methods are preferred when they achieve *higher* ROUGE-2 at *lower* latency. Compared with autoregressive decoding and standard speculative decoding, the static ensemble traces a Pareto front that provides consistent improved time–quality tradeoffs. Notably, DIVERSED improves upon this frontier, achieving higher quality at comparable latency and/or lower latency at comparable quality, thereby outperforming the Pareto-optimal set induced by the static ensemble.

Table 2: Comparison of finetuned draft vs. DIVERSED at temperature 1. Fine-tuned draft model yields higher acceptance rates but lower task performance.

| **Train/Test Dataset** | **Method** | **CNNDM** Accept Rate (↑) | **CNNDM** ROUGE-2 (↑) |
|---|---|---|---|
| CNNDM | SD | 21.06% | 9.46 |
| CNNDM/CNNDM | DIVERSED | **69.96%** | **12.11** |
| GSM8K/CNNDM | DIVERSED | 51.07% | 8.63 |
| **(a)** Draft/Target: Llama-3.2-1B / Llama-3.1-8B | | | |
| CNNDM | SD | 35.26% | 9.85 |
| CNNDM/CNNDM | DIVERSED | **46.59%** | **10.97** |
| GSM8K/CNNDM | DIVERSED | **62.55%** | 8.13 |
| **(b)** Draft/Target: Qwen3-0.6B / Qwen3-8B | | | |
| | | **GSM8K** Accept Rate (↑) | **GSM8K** Accuracy (↑) |
| GSM8K | SD | 44.60% | 67% |
| GSM8K/GSM8K | DIVERSED | **72.61%** | **67%** |
| CNNDM/GSM8K | DIVERSED | 72.57% | 62% |
| **(c)** Draft/Target: Llama-3.2-1B / Llama-3.1-8B | | | |
| GSM8K | SD | 59.58% | **91%** |
| GSM8K/GSM8K | DIVERSED | **76.48%** | 88% |
| CNNDM/GSM8K | DIVERSED | 74.30% | 78% |
| **(d)** Draft/Target: Qwen3-0.6B / Qwen3-8B | | | |

### 6.3 Case Studies

**Benchmark Zero-shot Performance.** We train DIVERSED on GSM8K and test the trained model on CNNDM, and vice versa (i.e., DIVERSED trained on CNNDM and tested on GSM8K).

Compared with standard speculative decoding, DIVERSED can be trained on one dataset and evaluated on another (e.g., trained on GSM8K and tested on CNNDM, or vice versa). In these cross-dataset settings, we observe higher acceptance rates but degraded task performance. This is expected: when training and testing on the same dataset, DIVERSED can learn task-specific relaxation that better preserves solution quality. These results further suggest that, to maintain strong performance, relaxation of the rejection rule should be task-dependent rather than universal.

**Fine-tuned draft model *vs.* dynamic ensemble.** We fine-tune the draft model separately on each dataset and then use the fine-tuned draft within standard speculative decoding. Table 3 reports results on GSM8K, XSum, and MBPP. Overall, fine-tuning yields accuracy/pass@1 comparable to DIVERSED, but DIVERSED consistently achieves the highest acceptance rate. Importantly, fine-tuning does not reliably improve acceptance: for the Llama pair, acceptance typically increases after fine-tuning, whereas for Qwen3 it decreases in most settings. These findings suggest that acceptance in speculative decoding is gov-

Table 3: Comparison of finetuned draft vs. DIVERSED at two temperatures $T = 1.0, 0.1$. Columns are grouped by task. Each block compares SD, SD (Finetuned draft), and DIVERSED for a given *Draft/Target* pair. It shows fine-tuning does not reliably improve acceptance.

| | **Temperature** | **GSM8K** | | **XSum** | | **MBPP** | |
|---|---|---|---|---|---|---|---|
| | | Accept Rate (↑) | Accuracy (↑) | Accept Rate (↑) | ROUGE-2 (↑) | Accept Rate (↑) | Pass@1 (↑) |
| SD | 1.0 | 44.60% | 67% | 20.44% | 7.09 | 26.30% | 53% |
| SD (Finetuned draft) | | 66.44% | 67% | 35.41% | 6.95 | 47.07% | 53% |
| DIVERSED (Ours) | | **72.61%** | 67% | **70.53%** | **7.23** | **85.03%** | 53% |
| **(a)** Draft/Target model pair is Llama-3.2-1B / Llama-3.1-8B. | | | | | | | |
| SD | 1.0 | 59.58% | 91% | 15.60% | 4.95 | 58.72% | 55% |
| SD (Finetuned draft) | | 57.54% | 90% | 26.00% | 4.66 | 11.70% | 55% |
| DIVERSED (Ours) | | **76.48%** | 88% | **47.01%** | **5.01** | **81.46%** | **59%** |
| **(b)** Draft/Target model pair is Qwen3-0.6B / Qwen3-8B. | | | | | | | |
| SD | 0.1 | 66.40% | 76% | 17.69% | 7.82 | 72.94% | 61% |
| SD (Finetuned draft) | | 52.81% | 76% | 18.81% | **7.97** | 74.32% | 61% |
| DIVERSED (Ours) | | **89.14%** | 76% | **64.75%** | 7.52 | **89.78%** | **62%** |
| **(c)** Draft/Target model pair is Llama-3.2-1B / Llama-3.1-8B. | | | | | | | |
| SD | 0.1 | 58.83% | 92% | 16.45% | 5.26 | 83.85% | 67% |
| SD (Finetuned draft) | | 60.45% | 92% | 18.16% | 5.62 | 48.27% | 67% |
| DIVERSED (Ours) | | **90.26%** | 92% | **71.22%** | **7.84** | **91.23%** | 67% |
| **(d)** Draft/Target model pair is Qwen3-0.6B / Qwen3-8B. | | | | | | | |

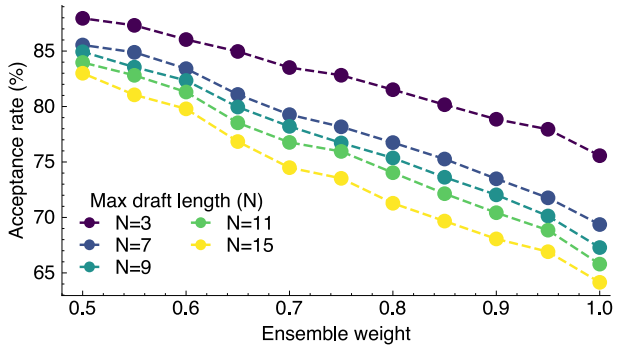


Figure 6: Acceptance rate decreases approximately linearly with the ensemble weight for static ensemble on CNNDM for the *Target/Draft* pair Llama-3.1-8B-Instruct/Llama-3.2-1B-Instruct at temperature 0.

erned primarily by *distributional alignment* between the draft and target conditionals, rather than by the draft model's standalone task performance.

**Impact of the static ensemble weight.** Figure 6 shows that the acceptance rate decreases as the ensemble weight $w$ increases. In the limit $w \to 1$, static-ensemble verification reduces to standard (lossless) speculative decoding. We report results on CNNDM using the Target/Draft pair Llama-3.1-8B-Instruct / Llama-3.2-1B-Instruct at temperature 0. This trend is consistent across draft lengths $N$ and aligns with Lemma 1, which predicts that acceptance decreases monotonically with $w$.

## 7 CONCLUSION

We studied lossy speculative decoding as a practical approach to accelerate large language model inference by increasing draft-token acceptance while preserving generation quality. Our empirical results show that optimal acceptance behavior is inherently context- and task-dependent, limiting the effectiveness of static verification rules.

To address this, we propose DIVERSED, a learned *dynamic* ensemble verifier that adapts to the local context at each decoding step and is trained in a task-specific manner. On the theory side, we derive an exact expression for the expected acceptance length without i.i.d. assumptions, yielding a more precise characterization than prior analyses. Empirically, we find that higher acceptance reliably translates into lower wall-clock latency, and DIVERSED consistently achieves higher acceptance at comparable or improved task quality.

Future work includes extending relaxed verification from token-level to block-level decisions and exploring cross-task transfer for the learned dynamic verifier.

**Acknowledgments** Ziyi Wang conducted this work during an Amazon internship. Nan Jiang is supported by TACC CCR25054.

## References


Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., et al. (2021). Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Bachmann, G., Anagnostidis, S., Pumarola, A., Georgopoulos, M., Sanakoyeu, A., Du, Y., Schönfeld, E., Thabet, A. K., and Kohler, J. (2025). Judge decoding: Faster speculative sampling requires going beyond model alignment. In *ICLR*. OpenReview.net.

Cai, T., Li, Y., Geng, Z., Peng, H., Lee, J. D., Chen, D., and Dao, T. (2024). Medusa: Simple llm inference acceleration framework with multiple decoding heads. In *ICML*, volume 235, pages 5209–5235.

Chen, C., Borgeaud, S., Irving, G., Lespiau, J.-B., Sifre, L., and Jumper, J. (2023). Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Dao, T., Fu, D., Ermon, S., Rudra, A., and Ré, C. (2022). Flashattention: Fast and memory-efficient exact attention with io-awareness. *NeurIPS*, 35:16344–16359.

Dettmers, T., Lewis, M., Belkada, Y., and Zettlemoyer, L. (2022). Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *NeurIPS*, 35:30318–30332.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Elhoushi, M., Shrivastava, A., Liskovich, D., Hosmer, B., Wasti, B., Lai, L., Mahmoud, A., Acun, B., Agarwal, S., Roman, A., Aly, A. A., Chen, B., and Wu, C. (2024). Layerskip: Enabling early exit inference and self-speculative decoding. In *ACL*, pages 12622–12642. Association for Computational Linguistics.

Goel, R., Gagrani, M., Jeon, W., Park, J., Lee, M., and Lott, C. (2024). Direct alignment of draft model for speculative decoding with chat-fine-tuned LLMs. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Hu, J., Liu, J. K., Xu, H., and Shen, W. (2025). Reinforce++: Stabilizing critic-free policy optimization with global advantage normalization.

Huang, K., Guo, X., and Wang, M. (2025). Specdec++: Boosting speculative decoding via adaptive candidate lengths. In *Second Conference on Language Modeling*.

Jiang, J., Wang, F., Shen, J., Kim, S., and Kim, S. (2024). A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In *NeurIPS*, volume 35, pages 22199–22213.

Leviathan, Y., Kalman, M., and Matias, Y. (2023). Fast inference from transformers via speculative decoding. In *ICML*, pages 19274–19286. PMLR.

Li, M., Chen, X., Holtzman, A., Chen, B., Lin, J., Yih, S., and Lin, V. (2024a). Nearest neighbor speculative decoding for LLM generation and attribution. In *NeurIPS*, volume 37, pages 80987–81015.

Li, Y., Wei, F., Zhang, C., and Zhang, H. (2024b). EAGLE: speculative sampling requires rethinking feature uncertainty. In *ICML*, volume 235, pages 28935–28948. OpenReview.net.

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

Liu, X., Hu, L., Bailis, P., Cheung, A., Deng, Z., Stoica, I., and Zhang, H. (2024). Online speculative decoding. In *ICML*, volume 235, pages 31131–31146. OpenReview.net.

Mamou, J., Pereg, O., Korat, D., Berchansky, M., Timor, N., Wasserblat, M., and Schwartz, R. (2024). Dynamic speculation lookahead accelerates speculative decoding of large language models. In *NeurIPS Efficient Natural Language and Speech Processing Workshop*, pages 456–467. PMLR.

Miao, X., Oliaro, G., Zhang, Z., Cheng, X., Wang, Z., Zhang, Z., Wong, R. Y. Y., Zhu, A., Yang, L., Shi, X., Shi, C., Chen, Z., Arfeen, D., Abhyankar, R., and Jia, Z. (2024). Specinfer: Accelerating large language model serving with tree-based speculative inference and verification. In *ASPLOS*, page 932–949, New York, NY, USA. Association for Computing Machinery.

Narasimhan, H., Jitkrittum, W., Rawat, A. S., Kim, S., Gupta, N., Menon, A. K., and Kumar, S. (2025). Faster cascades via speculative decoding. In *ICLR*. OpenReview.net.

Narayan, S., Cohen, S. B., and Lapata, M. (2018). Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *EMNLP*, pages 1797–1807. Association for Computational Linguistics.

Ni, Y., Liu, C., Tang, Y., Han, K., and Wang, Y. (2025). EMS-SD: Efficient multi-sample speculative decoding for accelerating large language models. In *NAACL*, pages 9307–9320. Association for Computational Linguistics.

Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., Zhao, S., Hong, L., Tian, R., Xie, R., Zhou, J., Gerstein, M., Li, D., Liu, Z., and Sun, M. (2024). Toolllm: Facilitating large language models to master 16000+ real-world apis. In *ICLR*. OpenReview.net.

Rozière, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X. E., Adi, Y., Liu, J., Sauvestre, R., Remez, T., Rapin, J., Kozhevnikov, A., Evtimov, I., Bitton, J., Bhatt, M., Ferrer, C. C., Grattafiori, A., Xiong, W., Défossez, A., Copet, J., Azhar, F., Touvron, H., Martin, L., Usunier, N., Scialom, T., and Synnaeve, G. (2023). Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., and Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. In *NeurIPS*, volume 36, pages 68539–68551.

See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In *ACL*, pages 1073–1083. Association for Computational Linguistics.

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al. (2025). Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., brian ichter, Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, volume 35, pages 24824–24837.

Yan, M., Agarwal, S., and Venkataraman, S. (2025). Decoding speculative decoding. In *NAACL (Long Papers)*, pages 6460–6473. Association for Computational Linguistics.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yin, M., Chen, M., Huang, K., and Wang, M. (2024). A theoretical perspective for speculative decoding algorithm. In *NeurIPS*, volume 37, pages 128082–128117.

Zhang, H. and Zhao, H. (2018). Minimum divergence vs. maximum margin: an empirical comparison on seq2seq models. In *ICLR*. OpenReview.net.

Zhang, J., Wang, J., Li, H., Shou, L., Chen, K., Chen, G., and Mehrotra, S. (2024). Draft& verify: Lossless large language model acceleration via self-speculative decoding. In *ACL*, pages 11263–11282. Association for Computational Linguistics.

Zhou, Y., Lyu, K., Rawat, A. S., Menon, A. K., Rostamizadeh, A., Kumar, S., Kagy, J., and Agarwal, R. (2024). Distillspec: Improving speculative decoding via knowledge distillation. In *ICLR*. OpenReview.net.

Zou, J., Roy, S., Verma, V. K., Wang, Z., Wipf, D., Lu, P., Negi, S., Zou, J., and He, J. (2025a). Tattoo: Tool-grounded thinking PRM for test-time scaling in tabular reasoning. *CoRR*, abs/2510.06217.

Zou, J., Yang, X., Qiu, R., Li, G., Tieu, K., Lu, P., Shen, K., Tong, H., Choi, Y., He, J., Zou, J., Wang, M., and Yang, L. (2025b). Latent collaboration in multi-agent systems. *CoRR*, abs/2511.20639.

## Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [No]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

# Supplementary Materials: DIVERSED: Relaxed Speculative Decoding via Dynamic Ensemble Verification

## A Training and Inference Pipeline of DIVERSED

**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 $x_{1:T}$.

$n \leftarrow 1$;
**while** $n \leq T$ **do**
  **for** $t = n$ **to** $n + N$ **do**
    sample $\tilde{x}_t \sim q_t(\cdot|x_{0:n-1}, \tilde{x}_{n:t-1})$ ▷ draft rollout
    cache hidden states $h^q_{n:n+N}$ and distribution $q_{n:n+N}$.
  **for** $t = n$ **to** $n + N$ **do**
    compute $p_t(\cdot|x_{0:n-1}, \tilde{x}_{n:t-1})$, and cache hidden states $h^p_{n:n+N}$. ▷ target verification
  compute the relaxed accept rule $(b_{\nu_t^\theta}, P_{\nu_t^\theta})$ in equation (6). ▷ dynamic ensemble in Section 3.2
  **for** $t = n$ **to** $n + N$ **do**
    Draw $u \sim \text{Uniform}(0, 1)$.
    **if** $u \leq b_{\nu_t^\theta}(\tilde{x}_t|x_{1:n-1}, \tilde{x}_{n:t-1})$ **then**
      **accept** $\tilde{x}_t$;
      set $x_n \leftarrow \tilde{x}_t$;
      update $n \leftarrow n + 1$.
    **else**
      **reject** $\tilde{x}_t$; ▷ discard $\tilde{x}_{t+1:T}$ and resume
      sample $x_n \sim P_{\nu_t^\theta}$;
      update $n \leftarrow n + 1$;
      **break**

# B Theoretical Justification

Table 4: List of notations and their definitions.

| Notation | Definition |
|---|---|
| $x_0$ | Input prompt |
| $N$ | Draft sequence length |
| $x_{1:N}$ | 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 |
| $q(\cdot\|x_{1:t-1})$ or $q_t(\cdot)$ | Draft (small) model sampling distribution at time $t$. |
| $p(\cdot\|x_{1:t-1})$ or $p_t(\cdot)$ | Target (small) model sampling distribution at time $t$. |
| $\nu$ | static ensemble for between draft and target model. |
| $\nu^\theta$ | Proposed dynamic ensemble for between draft and target model. |

## B.1 Proof of Proposition 1

*Proof.* Recall that $\nu_t$ is defined in Equation (2) as the static ensemble

$$\nu_t(\cdot) = w\, p_t(\cdot) + (1-w)\, q_t(\cdot).$$

The rejection probability under the acceptance rule $b_{\nu_t}(x) = \min\{1, \nu_t(x)/q_t(x)\}$ is

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

Next, the total variation distance between $\nu$ and $p$ satisfies:

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

Combining the two identities gives

$$\mathbb{P}_{\nu_t}(\text{reject}) + \mathrm{TV}(\nu_t, p_t) \;=\; \mathrm{TV}(q_t, p_t),$$

which concludes the proof. □

## B.2 Proof of Lemma 1

*Proof.* Conditioned on $x_{0:t-1}$, the draft proposal $\tilde{x}_t$ is sampled from $q_t$. The expected probability of accepting the $\tilde{x}_t$ given $x_{0:t-1}$ is:

$$\mathbb{E}_{x\sim q_t}\left[\min\left\{1, \frac{\nu_t(x)}{q_t(x)}\right\}\right] = \sum_x \min\left\{1, \frac{\nu_t(x)}{q_t(x)}\right\} q_t(x) = \sum_x \min\left\{q_t(x), \nu_t(x)\right\}$$

Using $\nu_t(x) = wp_t(x) + (1-w)q_t(x)$ as defined in Equation (2), we obtain

$$\sum_x \min\{q_t(x), \nu_t(x)\} = \sum_x \min\{q_t(x), wp_t(x) + (1-w)q_t(x)\} = \sum_x q_t(x) + \sum_x \min\{0, w(p_t(x) - q_t(x))\}$$
$$= 1 - w + \sum_x w \min\{q_t(x), p_t(x)\}$$

Finally, recall the identity $\mathrm{TV}(q_t, p_t) = 1 - \sum_x \min\{q_t(x), p_t(x)\}$. Substituting yields

$$(1-w) + w\sum_x \min\{q_t(x), p_t(x)\} = 1 - w\Big(1 - \sum_x \min\{q_t(x), p_t(x)\}\Big) = 1 - w\ \mathrm{TV}(q_t, p_t),$$

as claimed. □

### B.3 Proof of Lemma 2

*Proof.* Condition on the prefix $x_{0:t-1}$. By Bayes' rule,

$$\mathbb{P}(x_t = x | x_t \text{ is accepted}, x_{0:t-1}) = \frac{\mathbb{P}(x_t \text{ is accepted} | x_t = x, x_{0:t-1})\,\mathbb{P}(x_t = x \mid x_{0:t-1})}{\sum_{x'} \mathbb{P}(x_t \text{ is accepted} | x_t = x', x_{0:t-1})\,\mathbb{P}(x_t = x' \mid x_{0:t-1})}.$$

Under lossless speculative decoding, a draft proposal $x$ is accepted with probability $\min\{1, \frac{p(x|x_{0:t-1})}{q(x|x_{0:t-1})}\}$, and $\mathbb{P}(x_t = x \mid x_{0:t-1}) = q(x|x_{0:t-1})$. Substituting these expressions yields

$$\mathbb{P}(x_t = x | x_t \text{ is accepted}, x_{0:t-1}) = \frac{\min\Big\{1, \frac{p(x|x_{0:t-1})}{q(x|x_{0:t-1})}\Big\} q(x|x_{0:t-1})}{\sum_{x'} \min\Big\{1, \frac{p(x'|x_{0:t-1})}{q(x'|x_{0:t-1})}\Big\} q(x'|x_{0:t-1})} = \frac{\min\{q(x|x_{0:t-1}),\, p(x|x_{0:t-1})\}}{\sum_{x'} \min\{q(x'|x_{0:t-1}),\, p(x'|x_{0:t-1})\}}.$$

Finally, using a standard property of total variation distance $\sum_{x'} \min\{q(x'|x_{0:t-1}), p(x'|x_{0:t-1})\} = 1 - \mathrm{TV}(q(\cdot|x_{0:t-1}), p(\cdot|x_{0:t-1}))$, we have:

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

This completes the proof. □

### B.4 Proof of Theorem 2

*Proof.* Fix the draft length $N$, recall that $A_i$ denote the *acceptance* indicator at step $i$. $A_i = 1$ if the $i$-th draft token $x_i$ is accepted given the prefix $x_{0:i-1}$, and $A_i = 0$ otherwise.

We start by computing the conditional expectation, given a draft generation $x_{1:N}$, the expected the acceptance length $T_N$ is

$$\mathbb{E}[T_N | x_{1:N}] = \sum_{t=1}^{N} \mathbb{P}(T_N \geq t | x_{1:N})$$

where the expectation of an indicator is the probability of its event. The event $\{T_N \geq t\}$ means the first $t$ draft tokens are accepted, i.e., $\bigcap_{i=1}^{t}\{A_i = 1\}$. Since the acceptance decisions across steps are *independent* given the prefixes, the probability can be factorized as:

$$\mathbb{P}(T_N \geq t | x_{1:N}) = \prod_{i=1}^{t} \mathbb{P}(A_i = 1 | x_{1:N}) = \prod_{i=1}^{t} \min\left\{1, \frac{p(x_i \mid x_{0:i-1})}{q(x_i \mid x_{0:i-1})}\right\} \tag{7}$$

where the acceptance probability $\mathbb{P}(A_i = 1 | x_{1:N})$ equals to the accept rule $b(x_i)$ under speculative decoding.

Taking the expectation over the draft sequence and using the tower law gives

$$\begin{aligned}
\mathbb{E}[T_N] = \mathbb{E}_{x_{1:N}\sim q}\left[\mathbb{E}[T_N|x_{1:N}]\right] &= \sum_{x_{1:N}} \left[\sum_{t=1}^{N}\prod_{i=1}^{t}\min\left\{1, \frac{p(x_i \mid x_{0:i-1})}{q(x_i \mid x_{0:i-1})}\right\}\right] q(x_{1:N}) \\
&= \sum_{x_{1:N}} \left[\sum_{t=1}^{N}\left[\prod_{i=1}^{t}\min\left\{1, \frac{p(x_i \mid x_{0:i-1})}{q(x_i \mid x_{0:i-1})}\right\}\right]\prod_{i=1}^{N} q(x_i|x_{0:i-1})\right] \\
&= \sum_{x_{1:N}} \left[\sum_{t=1}^{N}\prod_{i=1}^{t}\min\{q(x_i|x_{0:i-1}), p(x_i|x_{0:i-1})\}\prod_{i=t+1}^{N} q(x_i|x_{0:i-1})\right] \\
&= \sum_{t=1}^{N}\left[\sum_{x_{1:N}}\prod_{i=1}^{t}\min\{q(x_i|x_{0:i-1}), p(x_i|x_{0:i-1})\}\prod_{i=t+1}^{N} q(x_i|x_{0:i-1})\right] \\
&= \sum_{t=1}^{N}\underbrace{\sum_{x_{1:t}}\prod_{i=1}^{t}\min\{q(x_i|x_{0:i-1}), p(x_i|x_{0:i-1})\}}_{\text{denote this part as } S_t}. \qquad (8)
\end{aligned}$$

where we used $q(x_{1:N}) = \prod_{i=1}^{N} q(x_i \mid x_{0:i-1})$ and $\sum_{x_{t+1:N}}\prod_{i=t+1}^{N} q(x_i \mid x_{0:i-1}) = 1$.

We expand $S_t$ by summing out the last token $x_t$ first. Since the product separates into a factor that depends on $x_t$ and a factor that does not, we have

$$S_t = \sum_{x_{1:t-1}}\left(\prod_{i=1}^{t-1}\min\{q(x_i \mid x_{0:i-1}), p(x_i \mid x_{0:i-1})\}\right)\sum_{x_t}\min\{q(x_t \mid x_{0:t-1}), p(x_t \mid x_{0:t-1})\}.$$

Next, we use a identity of TV distance for discrete distributions $\sum_x \min\{q(x), p(x)\} = 1 - \mathrm{TV}(q, p)$, applied conditionally given the prefix $x_{0:t-1}$. This yields

$$S_t = \sum_{x_{1:t-1}}\left(\prod_{i=1}^{t-1}\min\{q(x_i \mid x_{0:i-1}), p(x_i \mid x_{0:i-1})\}\right)\Big(1 - \mathrm{TV}(q(\cdot \mid x_{0:t-1}), p(\cdot \mid x_{0:t-1}))\Big). \qquad (9)$$

Now we convert the remaining min-products into expectations under the min-kernel $\tilde{m}$ from Lemma 2. We expand the defined definition of $\tilde{m}$ as:

$$\min\{q(x_i \mid x_{0:i-1}), p(x_i \mid x_{0:i-1})\} = \Big(1 - \mathrm{TV}(q(\cdot \mid x_{0:i-1}), p(\cdot \mid x_{0:i-1}))\Big)\,\tilde{m}(x_i \mid x_{0:i-1}).$$

Substituting this identity into the product in (9) for each $i = 1, \ldots, t-1$ gives

$$\begin{aligned}
S_t &= \sum_{x_{1:t-1}}\left(\prod_{i=1}^{t-1}\Big(1 - \mathrm{TV}(q(\cdot \mid x_{0:i-1}), p(\cdot \mid x_{0:i-1}))\Big)\,\tilde{m}(x_i \mid x_{0:i-1})\right)\Big(1 - \mathrm{TV}(q(\cdot \mid x_{0:t-1}), p(\cdot \mid x_{0:t-1}))\Big) \\
&= \sum_{x_{1:t-1}}\left(\prod_{i=1}^{t-1}\tilde{m}(x_i \mid x_{0:i-1})\right)\left(\prod_{i=0}^{t}\Big(1 - \mathrm{TV}(q(\cdot \mid x_{0:i}), p(\cdot \mid x_{0:i}))\Big)\right),
\end{aligned}$$

Finally, the sum over $x_{1:t-1}$ is exactly expectation with respect to the sequential measure induced by $\tilde{m}$, hence

$$S_t = \mathbb{E}_{x_{1:t-1}\sim\tilde{m}}\left[\prod_{i=0}^{t}\Big(1 - \mathrm{TV}(q(\cdot \mid x_{0:i}), p(\cdot \mid x_{0:i}))\Big)\right].$$

Combined with Equation (8), we finish the proof. □

# C Extended Experimental Analysis

## C.1 Implementation Details

**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.

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× A100-40GB GPUs.

Table 5: Hyper-parameter configuration for each benchmark dataset.

| | CNNDM | XSUM | MBPP | GSM8K |
|---|---|---|---|---|
| Generation budget | 128 | 128 | 384 | 512 |
| Hyper Parameter $\lambda$ | 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 $\lambda$ | 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 $\lambda$ | 0.5 | 0.5 | 0.3 | 0.35 |
| Draft length $N$ | | | $\{3, 5, 7\}$ | |
| Rollout Number | | | 4 | |
| Learning Rate | | | 5e-4 | |
| Training Epochs | | | 2 | |

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

In terms of baseline selection, we omit Medusa (Cai et al., 2024) and EAGLE (Li et al., 2024b) because they modify either the target LLM or the draft LLM, whereas DIVERSED keeps both models unchanged. A direct comparison would therefore conflate improvements from verification with changes in model architecture, making the comparison less controlled and less fair.

We view DIVERSED as complementary to DISCO, Medusa, and EAGLE. In principle, DIVERSED could be layered on top of these approaches by replacing their rigid verification step with our relaxed, ensemble-based verifier, potentially yielding additional efficiency gains. However, due to engineering complexity and the limited rebuttal timeline, we did not evaluate these combinations. All baselines we report follow the same line of work that holds the draft and target models fixed and differs only in how the acceptance rule is defined.

### C.2 Extended Experiment Results

Table 6: 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 (↑) | Quality Accuracy (↑) | Accept Rate(↑) | Quality ROUGE-2 (↑) | Accept Rate(↑) | Quality ROUGE-2 (↑) | Accept Rate(↑) | Quality pass@1 (↑) |
| 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.

Table 7: 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.

| | **GSM8K** | | **CNNDM** | | **XSum** | | **MBPP** | |
|---|---|---|---|---|---|---|---|---|
| **Method** | Accept Rate (↑) | Quality Accuracy (↑) | 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.

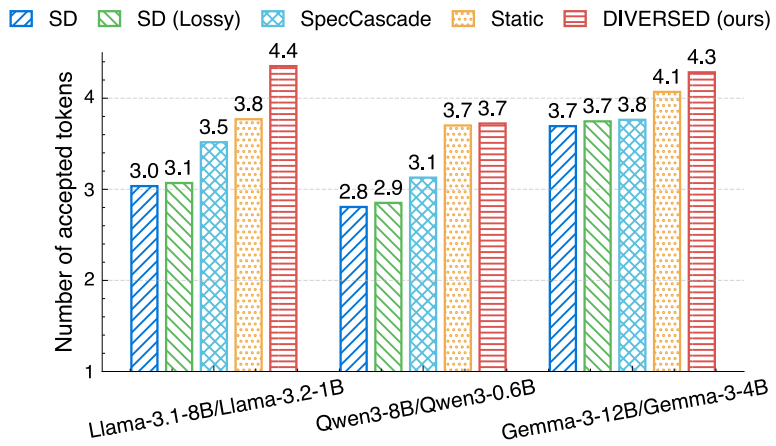


Figure 7: Average number of tokens accepted per verification round (temperature = 0.1, draft length $N = 5$) across three target/draft model pairs. Our DIVERSED consistently surpasses all baselines by achieving a higher number of accepted tokens.

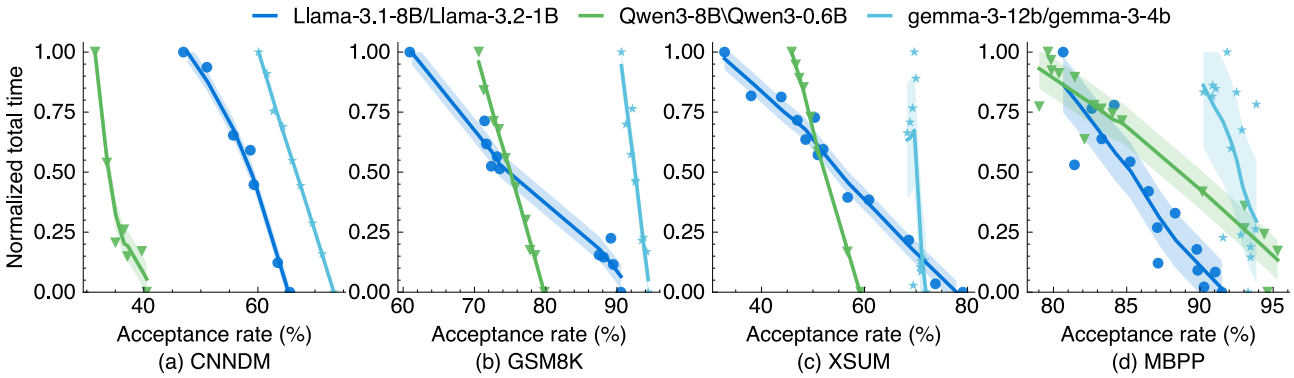


Figure 8: Results on the CNNDM, XSUM, MBPP, and GSM8k datasets for three target/draft pairs: normalized total time (lower is better) versus acceptance rate (higher is better) under speculative decoding with $N$=3. Model pairs with higher acceptance achieve lower decoding time.

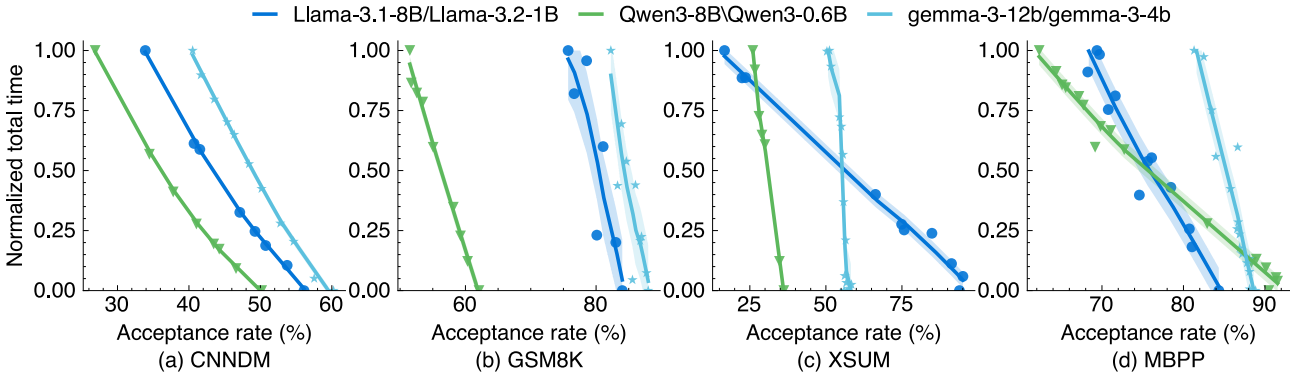


Figure 9: Results on the CNNDM, XSUM, MBPP, and GSM8k datasets for three target/draft pairs: normalized total time (lower is better) versus acceptance rate (higher is better) under speculative decoding with $N$=7. Model pairs with higher acceptance achieve lower decoding time.

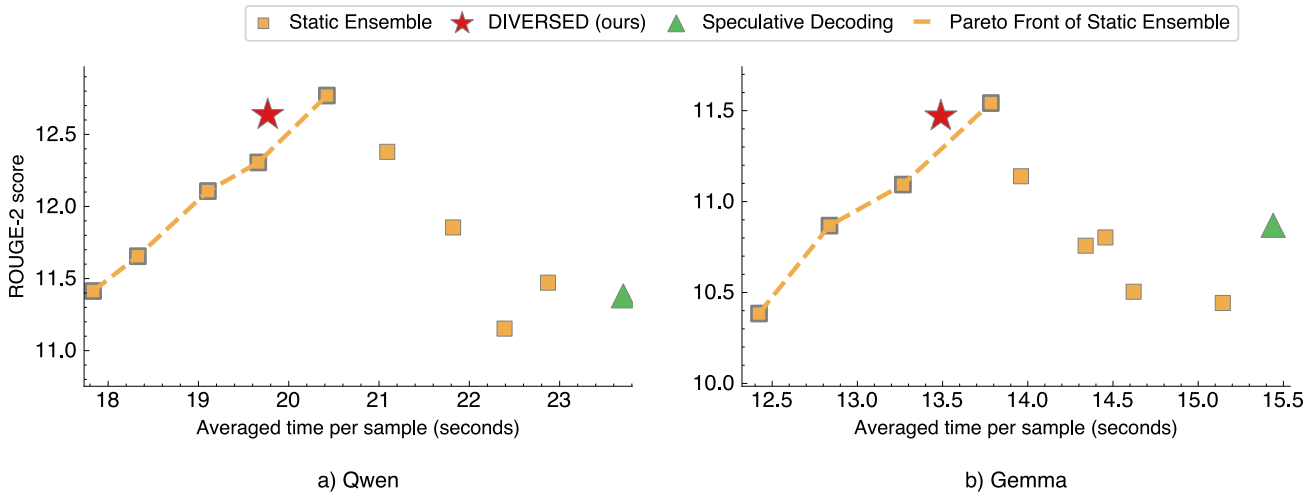


Figure 10: Our DIVERSED outperforms the Pareto front of the static ensemble and other baselines. Results are reported on the CNNDM dataset using the target/draft model pair: Qwen3-8B/Qwen3-0.6B and Gemma-3-12b-it/Gemma-3-4b-it, with temperature set to 0.

### C.3 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.<br>So in total she will have spent 1.5 + 1 + 1 = 3.5 hours walking the 12-mile trail.<br>###6 |
| SD | She has walked 4 + 2 = 6 miles in 2 hours.<br>So, she has 12 - 6 = 6 miles left to walk.<br>She wants to walk the entire trail in 12 / 4 = 3 hours.<br>So, she needs to walk the remaining 6 miles in 3 - 2 = 1 hour.<br>Therefore, she needs to walk the remaining 6 miles in 6 / 1 = 6 miles per hour.<br>###6 |
| SD (Lossy) | She has walked 4 + 2 = 6 miles in 2 hours.<br>So, she has 12 - 6 = 6 miles left to walk.<br>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.<br>Therefore, she needs to walk the remaining distance at a speed of 6 / 1.5 = 4 miles per hour.<br>###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.<br>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.<br>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.<br>###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.<br>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.<br>###3. |
| DIVERSED | She has walked 4 + 2 = 6 miles in 1 + 1 = 2 hours.<br>She has 12 - 6 = 6 miles left to walk.<br>She wants to walk the entire trail in 12 / 4 = 3 hours.<br>So she has 3 - 2 = 1 hour to walk the remaining 6 miles.<br>Therefore, she needs to walk 6 / 1 = 6 miles per hour.<br>###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 miles / 4 mph = 3 hours.<br>So far, she has spent 1 + 1 = 2 hours. This leaves her with 3 - 2 = 1 hour to walk the remaining distance.<br>She has 12 - 4 - 2 = 6 miles left. Therefore, she needs to walk 6 miles in 1 hour, which is a speed of 6 mph.<br>###6 |
| SD | The total distance of the trail is 12 miles. So far, she has walked 4 + 2 = 6 miles. Therefore, she has 12 - 6 = 6 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: total time = total distance / average speed = 12 miles / 4 mph = 3 hours. She has already spent 1 + 1 = 2 hours. Therefore, she has 3 - 2 = 1 hour left to walk the remaining 6 miles. To find the required speed, we can use the formula: speed = distance / time = 6 miles / 1 hour = 6 miles per hour. So, she needs to walk the remaining distance at 6 miles per hour.<br>###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 miles. She has 12 - 6 = 6 miles left. She has already spent 1 + 1 = 2 hours. To have an average speed of 4 miles per hour, she needs to complete the 12-mile trail in 12 / 4 = 3 hours. She has 3 - 2 = 1 hour left to walk the remaining 6 miles. Therefore, she needs to walk at a speed of 6.<br>###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 miles. She has 12 - 6 = 6 miles left. She has already spent 1 + 1 = 2 hours. To have an average speed of 4 miles per hour, the total time for the entire trail should be 12 / 4 = 3 hours. Therefore, she has 3 - 2 = 1 hour left to walk the remaining 6 miles. So, she needs to walk at a speed of 6.<br>###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 miles. She has 12 - 6 = 6 miles left. She has already spent 1 + 1 = 2 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 hours. Therefore, she has 3 - 2 = 1 hour left to walk the remaining 6 miles. So, she needs to walk at a speed of 6 / 1 = 6 miles per hour.<br>###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 hours. Therefore, she needs to walk the remaining 6 miles in 3 - 2 = 1 hour. So her speed for the remaining distance should be 6 / 1 = 6 miles per hour.<br>###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.<br>The total time she wants to take for the whole trail is 12 / 4 = 3 hours.<br>She has already spent 1 + 1 = 2 hours hiking, so she has 3 - 2 = 1 hour left.<br>The required speed for the last 6 miles in 1 hour is 6 / 1 = 6 miles per hour.<br>###6 |
| SD | Total distance is 12 miles, and she wants her average speed to be 4 miles per hour.<br>So, the total time she can spend hiking is 12/4 = 3 hours.<br>She has already spent 1 + 1 = 2 hours.<br>Therefore, she can hike the remaining distance for 3 - 2 = 1 hour.<br>The remaining distance is 12 - 4 - 2 = 6 miles.<br>So, she needs to walk 6 miles in 1 hour, which means she needs to walk at a speed of 6 miles per hour.<br>###6 |
| SD (Lossy) | The total distance of the trail is 12 miles.<br>Marissa has already walked 4 + 2 = 6 miles.<br>The remaining distance is 12 - 6 = 6 miles.<br>She wants her average speed to be 4 miles per hour for the entire 12-mile trail.<br>So, the total time she should take to complete the 12-mile trail is 12 / 4 = 3 hours.<br>She has already taken 1 + 1 = 2 hours.<br>The remaining time she has is 3 - 2 = 1 hour.<br>To walk the remaining 6 miles in 1 hour, she needs to walk at a speed of 6 / 1 = 6 miles per hour.<br>###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.<br>###6 |
| Static Ensemble | Total distance is 12 miles. Marissa has already walked 4 + 2 = 6 miles. The remaining distance is 12 - 6 = 6 miles.<br>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.<br>She has already walked for 1 + 1 = 2 hours. Therefore, she has 3 - 2 = 1 hour to walk the remaining 6 miles.<br>The speed she needs to walk the remaining distance is 6 miles / 1 hour = 6 miles per hour.<br>###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 miles / 4 miles per hour = 3 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 miles / 1 hour = 6 miles per hour.<br>###6 |

### C.4 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. |

Case Study on the CNNDM dataset using Qwen 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 | 1. President Obama highlights shrinking Florida Everglades as proof that climate change is real and a threat to national treasures and local economies .<br>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 .<br>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.<br>2. Obama criticizes climate change as a current issue, not a future threat.<br>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.<br>2. Obama's trip to Florida, which required 9,000 gallons of fuel for Air Force One, drew criticism over its environmental impact.<br>3. Obama emphasizes that climate change is a current issue with serious implications, not a problem for future generations.<br>4. Obama's administration is committed to leading the world in combating climate change through reducing carbon emissions.<br>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.<br>2. The president says climate change is not a problem for another generation.<br>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.<br>4. The White House said that climate change is affecting iconic places like the Everglades and Glacier National Park.<br>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.<br>2) Obama and Bill Nye 'the Science Guy' on trip.<br>3) Obama addresses climate change as a threat to national treasures and economies.<br>4) Obama mentions the impact of climate change on the Everglades and surrounding towns.<br>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.<br>2. Obama commits U.S. to lead in combatting climate change and reducing carbon emissions.<br>3. Obama's trip to Florida, accompanied by Bill Nye 'the Science Guy,' was seen as a challenge to his Earth Day message. |

Case Study on the CNNDM dataset using Gemma 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 Obama held up the shrinking Florida Everglades today as proof positive that climate change is real and is threatening America's 'national treasures'.<br>Obama says climate change is already having a negative effect on communities.<br>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 ẗhe Science Guyön 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.<br>2. Obama's trip to Florida included a walking tour of the Everglades and a speech on Earth Day.<br>3. The trip highlighted the administration's efforts to reduce carbon emissions.<br>4. Obama's gas-guzzling trip to Florida was criticized, but the White House defended it.<br>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.<br>2. Trip highlights shrinking Everglades and economic impact on surrounding towns.<br>3. Obama's Air Force One trip uses 9,000 gallons of fuel, raising questions about hypocrisy.<br>4. Bill Nye 'the Science Guy' accompanies Obama.<br>5. Obama emphasizes the urgency of addressing climate change and its impact on future generations." |

### C.5 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 ïllegally 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 ẗemporarily serve the sentence out of prisonäs she was s̈everely 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): Ï'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 ïllegally 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 ẗemporarily serve the sentence out of prisonäs she was s̈everely 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): Ï'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 ïllegally 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 ẗemporarily serve the sentence out of prisonäs she was s̈everely 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): Ï'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.6 MBPP Case Study

Case Study on the MBPP dataset using Llama model.

Write a function to find the median of three numbers.

| | |
|---|---|
| Autoregressive | `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]` |
| SD | `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]` |
| SD (Lossy) | `def median_numbers(a,b,c): return sorted([a,b,c])[1]` |
| SpecCascade | `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` |
| Static Ensemble | `def median_numbers(a,b,c): return sorted([a,b,c])[1]` |
| DIVERSED | `def median_numbers(a,b,c): return sorted([a,b,c])[1]` |

Case Study on the MBPP dataset using Qwen model.

Write a function to find the median of three numbers.

| | |
|---|---|
| Autoregressive | `def median_numbers(a,b,c): return sorted([a,b,c])[1]` |
| SD | `def median_numbers(a,b,c): return sorted([a,b,c])[1]` |
| SD (Lossy) | `def median_numbers(a,b,c): return sorted([a,b,c])[1]` |
| SpecCascade | `def median_numbers(a,b,c): return sorted([a,b,c])[1]` |
| Static Ensemble | `def median_numbers(a,b,c): return sorted([a,b,c])[1]` |
| DIVERSED | `def median_numbers(a,b,c): return sorted([a,b,c])[1]` |

Case Study on the MBPP dataset using Gemma model.

Write a function to find the median of three numbers.

| | |
|---|---|
| Autoregressive | `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` |
| SD | `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]` |
| SD (Lossy) | `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` |
| SpecCascade | `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` |
| Static Ensemble | `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` |
| DIVERSED | `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` |