Theoretical Benefit and Limitation of Diffusion Language Model

Guhao Feng*, Yihan Geng*, Jian Guan, Wei Wu, Liwei Wang, Di He

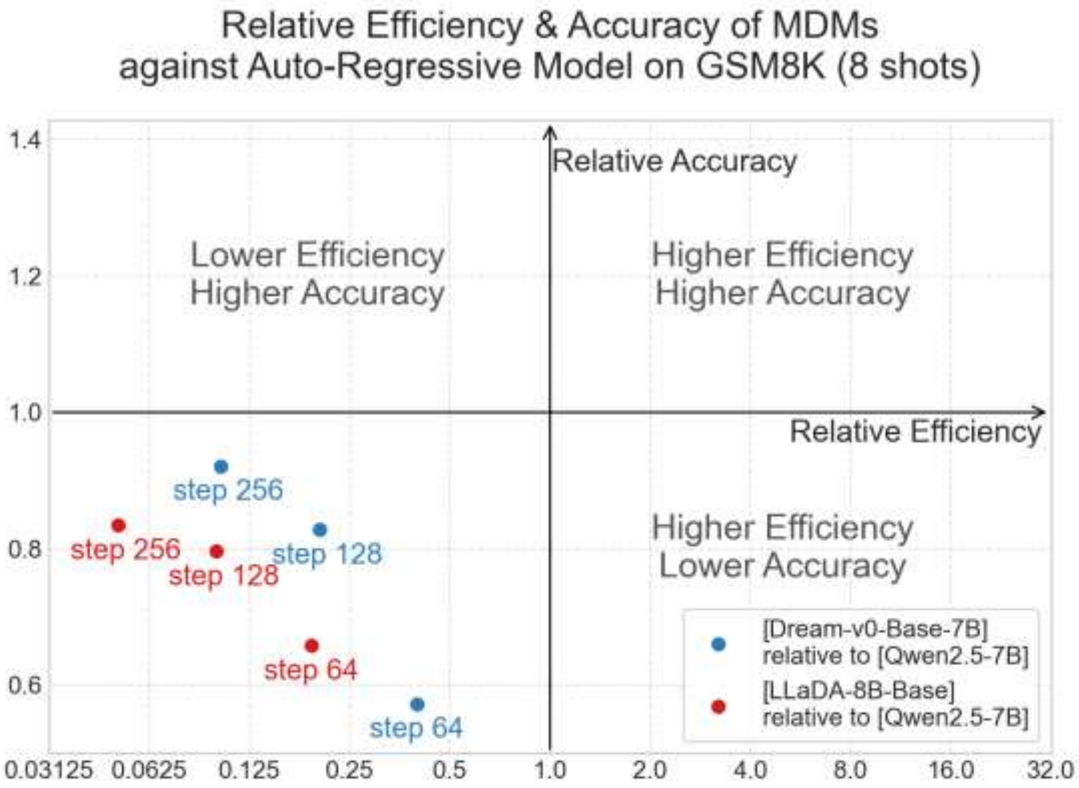
Paper: https://arxiv.org/abs/2502.09622

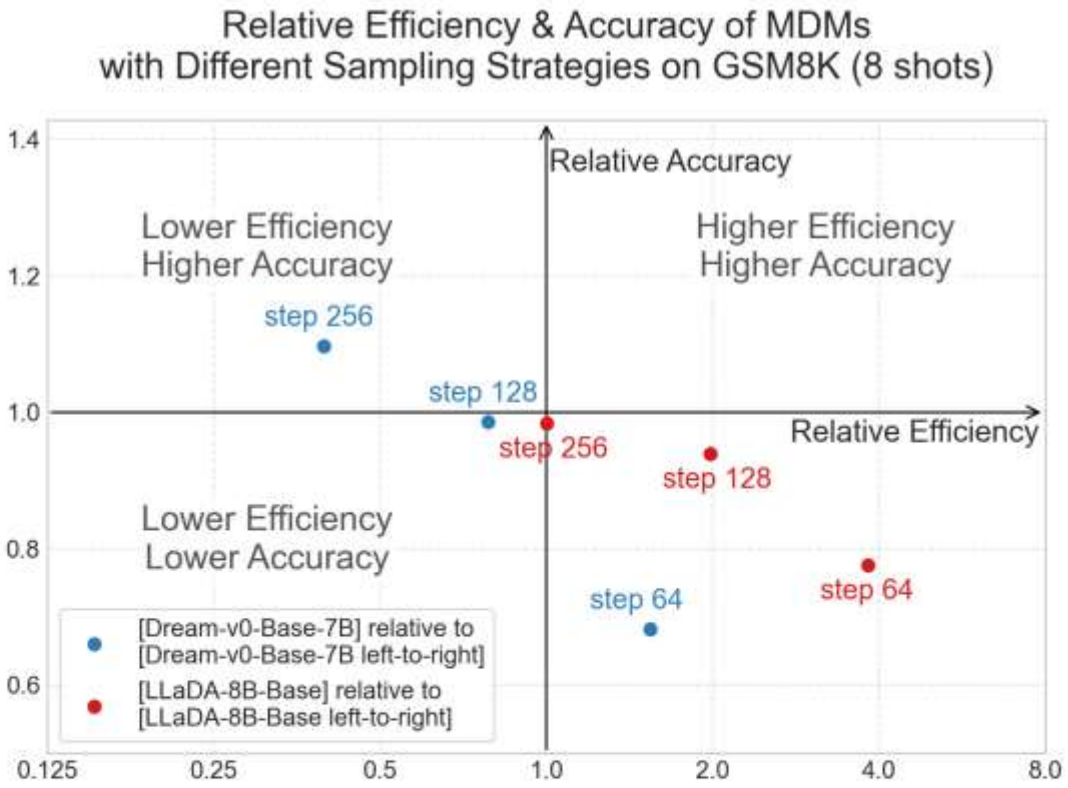
NeurIPS 2025 Poster

Background: Different LLMs

- Auto-regressive models (**ARs**): generate text token-by-token.
- Masked Diffusion Models (**MDMs**): at each step, generate **multiple** tokens independently.
  - Begins with an initial sequence of **masked tokens**, then iteratively **replaces masked tokens** with predicted ones in the sequence.
  - Intuition: parallel sampling can speed up inference
  - Previous studies: **low perplexity** with high efficiency

Empirical Observations on Reasoning Tasks





- Q: What is the trade-off between the accuracy and efficiency of MDMs?

* Length = 256

Our Evaluation Metrics: TER & SER

- **Token Error Rate (TER)**: Measures **fluency** via perplexity.
  - Ground-truth is q, and the evaluated model is p
  - $\mathrm{TER}(p) = 2^{\mathbb{E}_{\boldsymbol{x}\sim q}\left[\frac{-\log(p(\boldsymbol{x}))}{|\boldsymbol{x}|}\right]}$, where $\log(x)$ represents $\log_2(x)$
- **Sequence Error Rate (SER)**: Measures the **correctness** of the entire sequence, by evaluating the probability of generating a false sequence.
  - The target language q is defined on vocabulary $\mathcal{V}$
  - $\mathrm{SER}(p) = 1 - \sum_{\boldsymbol{x}\in\mathcal{L}_q} p(\boldsymbol{x})$, where $\mathcal{L}_q = \{\boldsymbol{x} \in \mathcal{V}^* \mid q(\boldsymbol{x}) > 0\}$ is the support set of q

Theoretical Analysis: TER

Theorem (TER, positive):

- Informally, MDMs achieve **near-optimal TER** with a **constant** number of steps for n-grams, approximately **independent** of sequence length.
- Intuition: MDMs can generate long sequences **efficiently with high fluency**.

Theorem 4.2 (TER Bounds for n-Gram Language Generation). *For any n-gram language q and any $\epsilon > 0$, let p_θ denote the reverse model and L denote the sequence length. The distribution over sequences generated by p_θ is denoted as p. For any $L > O(\frac{n-1}{\epsilon^{n+0.5}})$, under Assumption 4.1, there exists a masking schedule α_t such that, with $N = O(\frac{n-1}{\epsilon^n})$ sampling steps, the TER of the MDM is upper-bounded by:*

$$\log \mathrm{TER}(p) \leq \log \mathrm{TER}(q) + \epsilon_{learning} + 4\epsilon \log |\mathcal{V}|. \quad (6)$$

Theoretical Analysis: SER

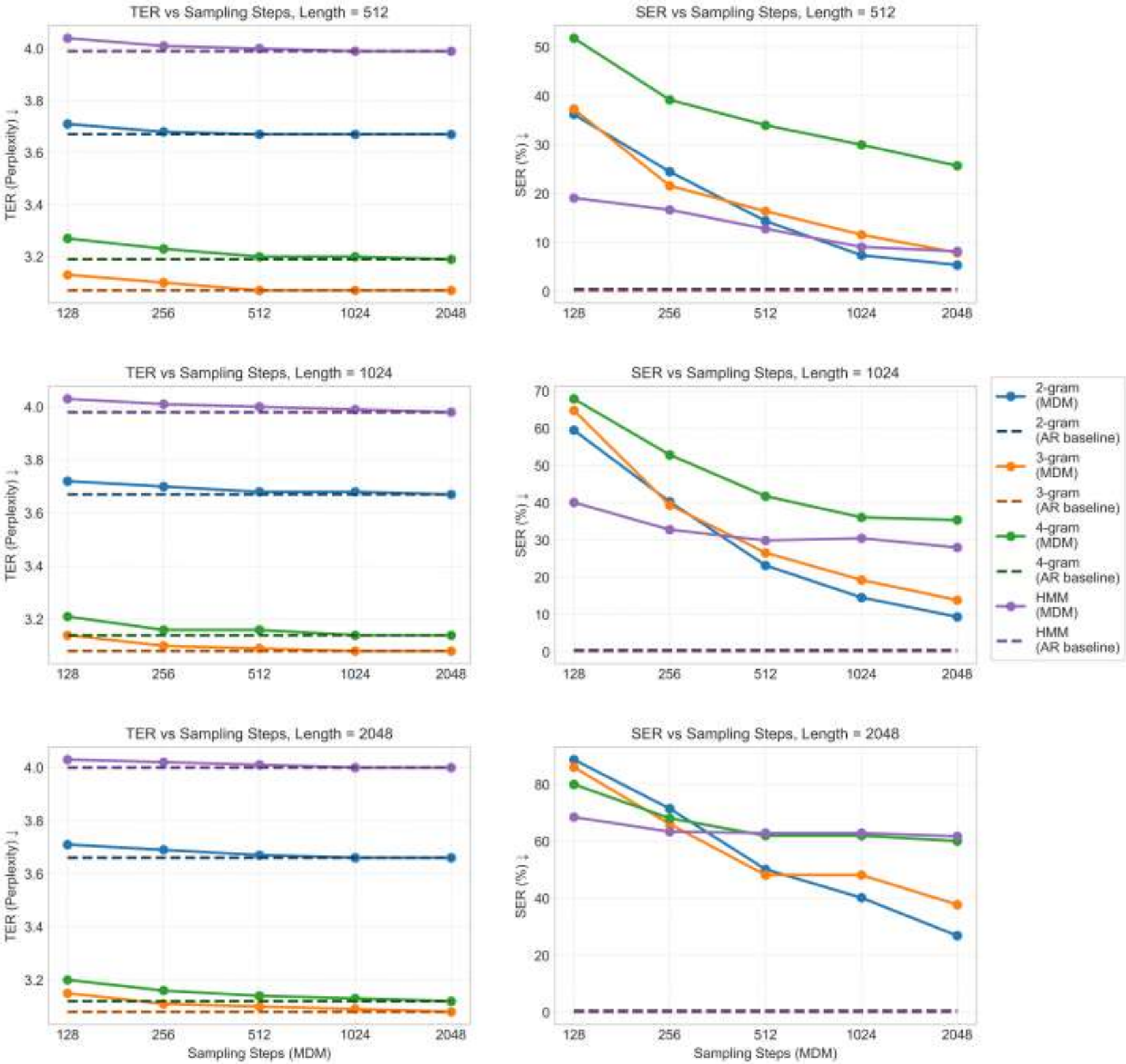
Theorem (SER, negative):

- Informally, even with steps **linear** to sequence length, there exists HMMs such that MDMs still have **a high SER**.
- Intuition: MDMs **do not offer a favorable trade-off** for tasks requiring **overall consistency and accuracy**.

Theorem 4.4 (SER Bound for HMM Generation)**.** *There exists an HMM q over a vocabulary of size 16 that satisfies the following conditions: for any reverse model p_θ under Assumption 4.1 with $\epsilon_{\text{learning}} < \frac{1}{128}$, and any masking schedule α_t, let p denote the distribution over sequences generated by p_θ. There exists a constant C such that if the number of sampling steps satisfies $N = CL$, where L is the sequence length, the SER of the generated text is lower-bounded by:* $\mathrm{SER}(p) > \frac{1}{2}$.

Experiments

- Setup: MDMs and AR baselines on formal languages.
- The results align with our theory:
- The TER of MDMs saturates after **fixed steps** (~512 steps) for all lengths.
- The SER of MDMs **decreases significantly more slowly** and depends on sequence length.



Inference Time Comparison

- With a **fixed** number of sampling steps (e.g., 512 steps), MDMs demonstrate **considerable efficiency** compared to ARs, especially for longer sequences.

Sequence Length	512	1024	2048
MDMs (512 steps)	3.1s	4.2s	4.7s
AR	1.7s	3.3s	7.0s

Conclusions and Take-aways

- MDMs can efficiently generate low-TER sentences, but may incur higher costs when evaluating the generation under SER.
- Efficiency is metric-sensitive: it depends on what we care about measuring.
- Practical guideline of ARs v.s. MDMs:
  - **MDMs** are better suited for **fluent generation tasks.**
  - **AR models** remain superior for tasks requiring **precise, step-by-step reasoning.**

Thanks!