Renoise Consistency: Unlocking Efficient Self-Correction for Diffusion Large Language Models

Jihoon Lee[1], Yeongbin Seo[2,3], Jaehyung Kim[1]
[1]Yonsei University [2]University of Toronto [3]Vector Institute
Machine & Language Learning Lab @ Yonsei

NEURAL INFORMATION PROCESSING SYSTEMS

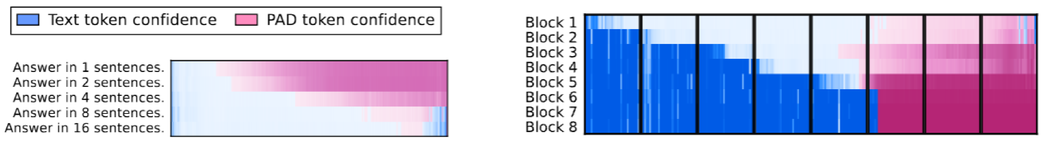
VECTOR INSTITUTE

Motivation

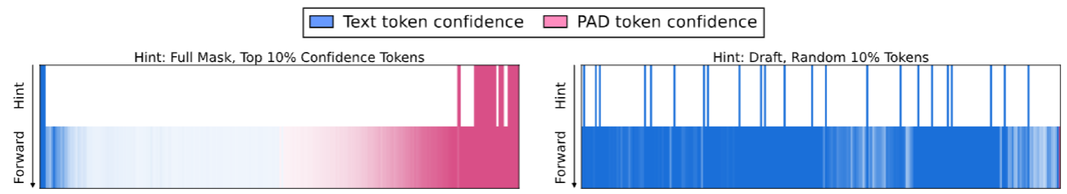
Draft	1000 - 356 = 664 1000 - 135 = 865 664 + 865 = 1529	1529	✗
Fail	⊠⊠⊠⊠ ⊠ ⊠⊠⊠ ⊠ ⊠64 ⊠⊠⊠0 ⊠ ⊠⊠⊠ ⊠ ⊠⊠⊠ ⊠⊠⊠ ⊠ ⊠⊠⊠ = ⊠⊠⊠⊠	664 → 1529	✗
Success	⊠⊠⊠⊠ - ⊠5⊠ ⊠ ⊠⊠⊠ ⊠⊠⊠⊠ ⊠ ⊠⊠⊠ = ⊠⊠⊠ ⊠⊠⊠ + ⊠⊠⊠ ⊠ ⊠⊠⊠⊠	644 → 1509	✓

Diffusion Large Language Models (DLMs) can self-correct.

Observation

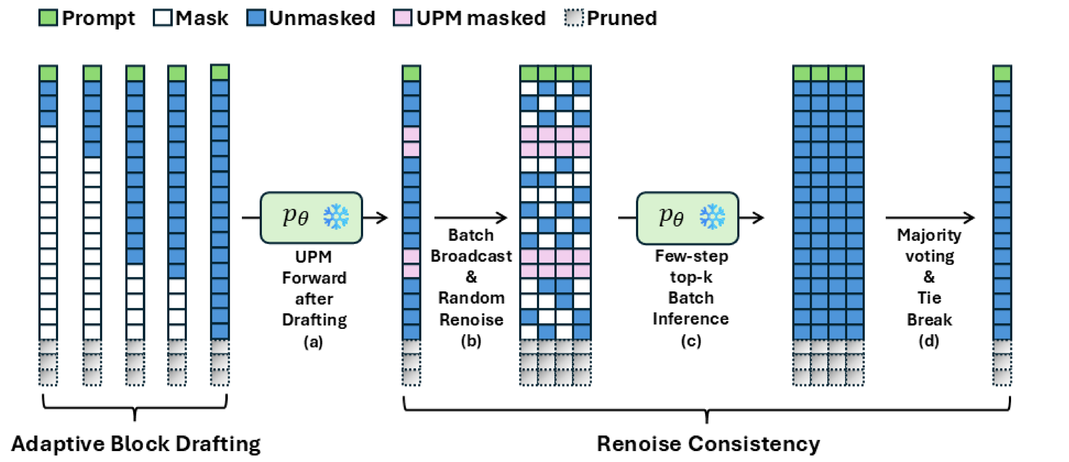


Findings #1: **Bidirectional Inertia** reveals the optimal **output length** or **block length.**
Adaptive **B**lock **D**rafting **(ABD),** efficient drafting.



Findings #2: **Scattered Anchor Context** resolves the **train-inference mismatch.**
Renoise **Con**sistency **(ReCon),** efficient self-correction.

Method

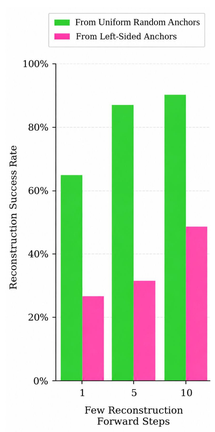


2 step train-free inference (ABD + ReCon)

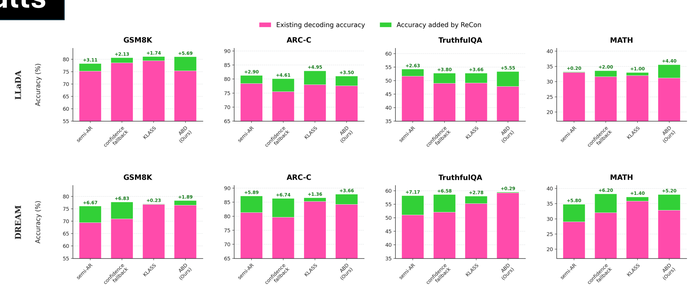
Takeaway

$$\mathcal{L}(\theta) \triangleq -\mathop{\mathbb{E}}_{t,x_0,x_t}\left[\frac{1}{t}\sum_{i=1}^{L}\mathbb{1}\left[x_t^i = M\right] log\, p_\theta\left(x_0^i \mid x_t\right)\right]$$

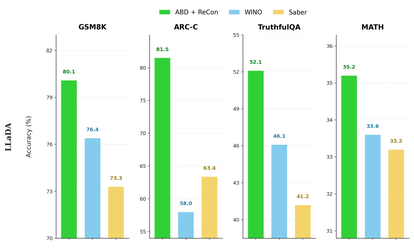
Uniform masking ($t \sim U[0,1]$) yields
scattered anchor contexts (train),
while existing decoding methods **(inference)**
impose a mismatched left-to-right prior.



Results



ReCon improves drafts.
Semi-AR: up to **7.17%p** | Conf. fallback: up to **6.83%p**
KLASS: up to **4.95%p** | ABD: up to **5.69%p**



ABD + ReCon outperforms
lightweight revocable decodings.
WINO: up to **23.5%p**
Saber: up to **18.1%p**