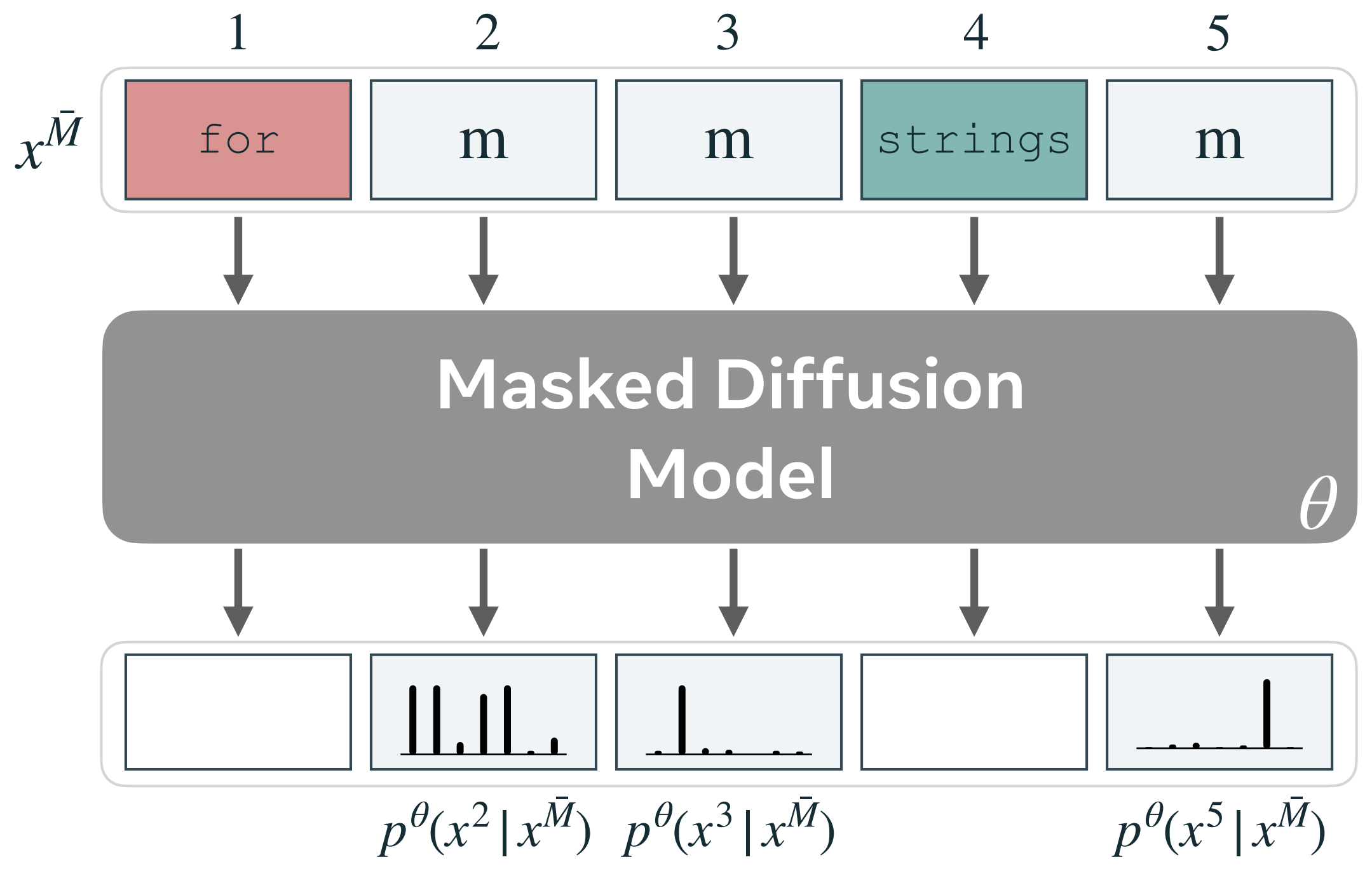
Accelerated Sampling from Masked Diffusion Models via Entropy Bounded Unmasking

Heli Ben-Hamu, Itai Gat, Daniel Severo, Niklas Nolte, Brian Karrer

FAIR, Meta AI

Motivation

Masked Diffusion Models (MDMs). Recent MDMs have shown competitive performance compared to autoregressive models (ARMs) for language modeling [1,2].


1
2
3
4
5
$x^{\bar{M}}$
for
m
m
strings
m
Masked Diffusion Model
θ
$p^\theta(x^2|x^{\bar{M}})$
$p^\theta(x^3|x^{\bar{M}})$
$p^\theta(x^5|x^{\bar{M}})$

- **Order of unmasking matters.** Random order sampling is inferior to greedy decoding.
- **Sampling efficiency.** MDMs hold promise for more efficient generation by simultaneously predicting multiple tokens, but are restricted by their factorized representation.

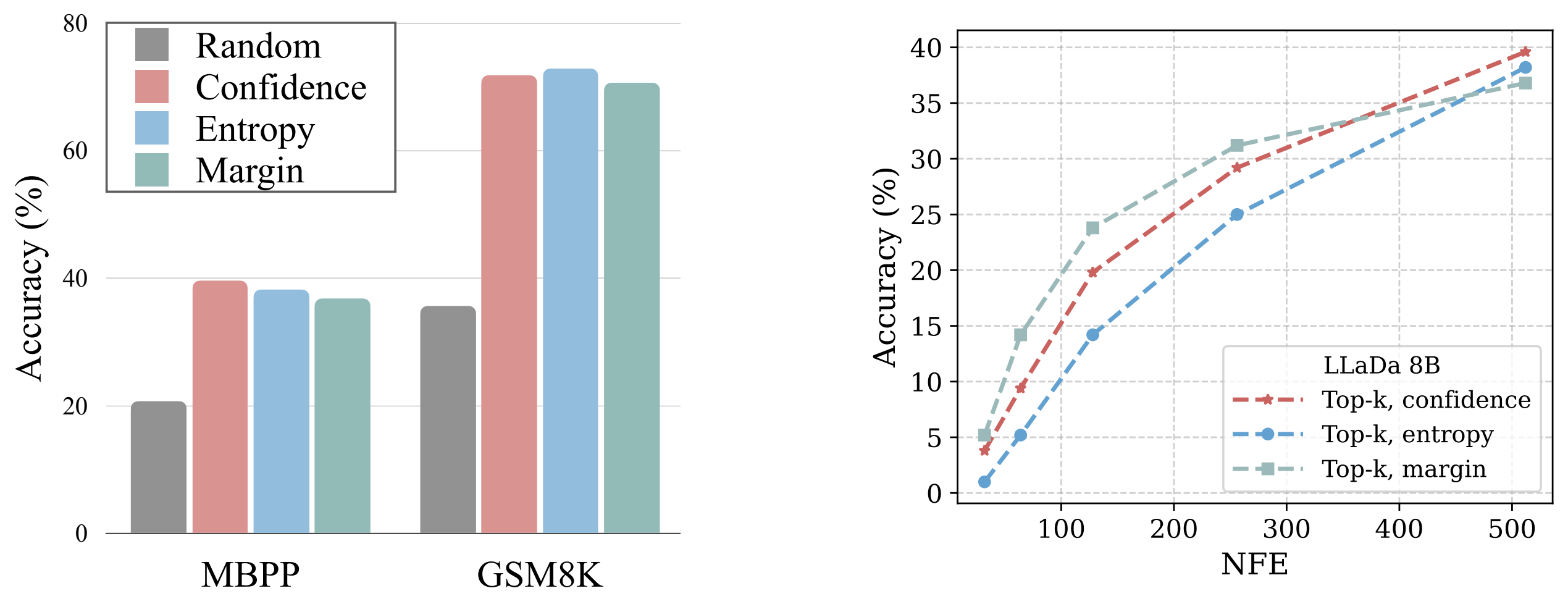


Figure 1. Left: pass@1 accuracy for various sampling strategies. Right: pass@1 accuracy on MBPP with Top-k sampling.

Error Decomposition

Let $q(x)$ denote the ground truth data distribution. Let a sampling procedure defined by a distribution over partitions of the sequence, denoted by ϕ. We use $p_\phi(x)$ to denote the model distribution with ϕ sampling. The KL divergence $D_{\mathrm{KL}}(q(x), p_\phi(x))$ can be bounded by a sum of two sources of error. At a certain step, for a subset of tokens, denoted by $\mathcal{S}$, sampled by ϕ:

$$\underbrace{\sum_{l\in\mathcal{S}} D_{\mathrm{KL}}\left(q(x^l|x^{\bar{M}}), p^\theta(x^l|x^{\bar{M}})\right)}_{\text{Model Error}} + \underbrace{D_{\mathrm{KL}}\left(q(x^{\mathcal{S}}|x^{\bar{M}}), \prod_{l\in\mathcal{S}} q(x^l|x^{\bar{M}})\right)}_{\text{Joint Dependence Error}}$$

When can we predict more than one token with zero error?

Key Observation

A **partially masked** token sequence often **deterministically sets multiple unknown tokens**. Thus, **standard sampling** procedures **do not fully use** the **information** present in a single masked model prediction.

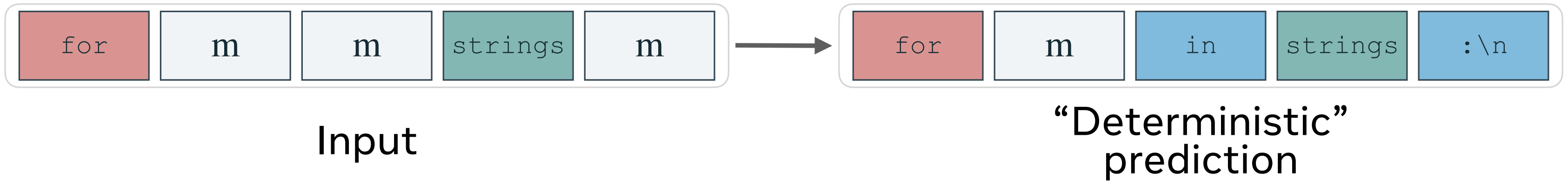


Figure 2. Example of a deterministic prediction scenario. When a `for` statement appears, it is (almost) certain that `in` and `:\n` will follow.

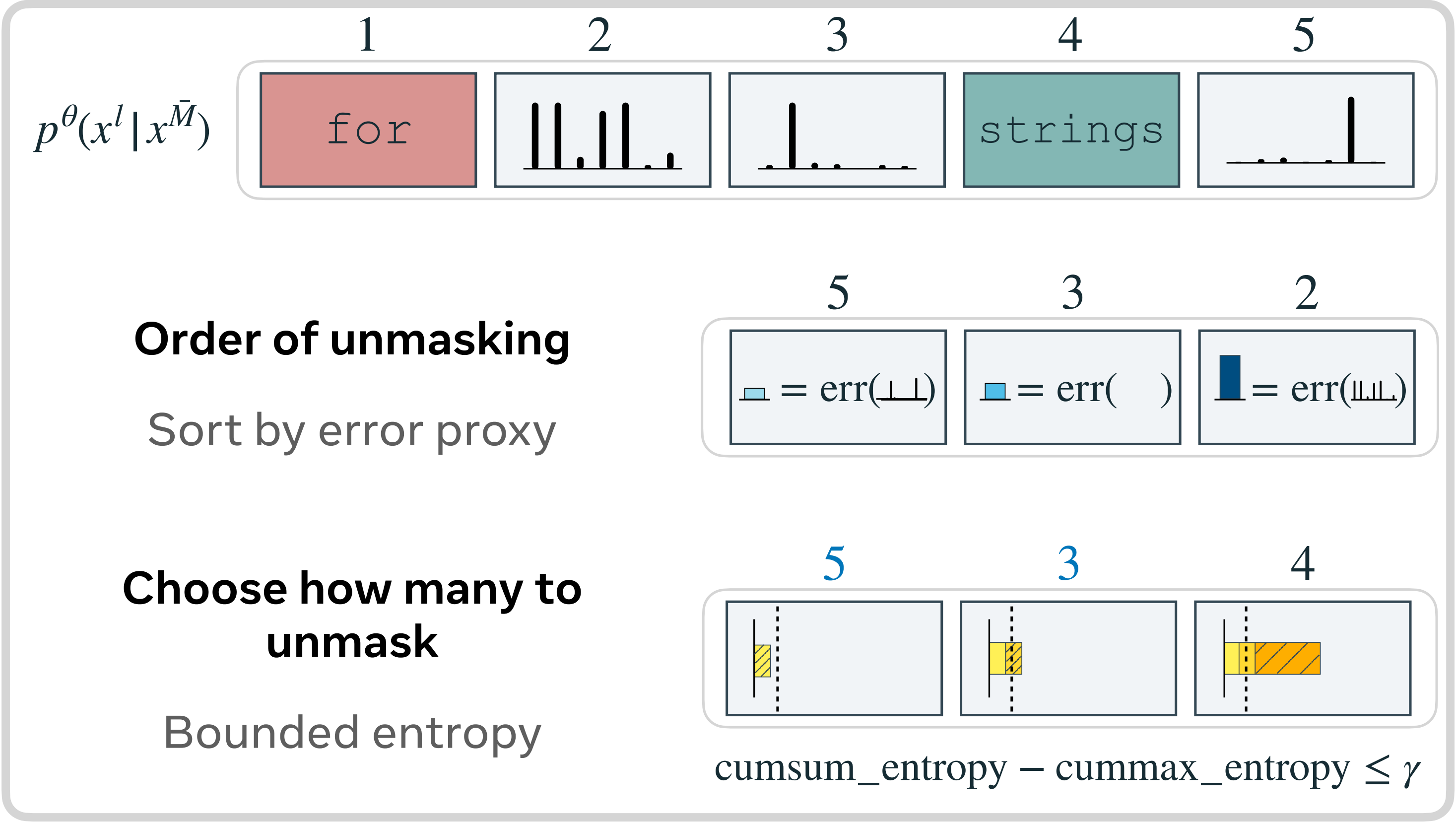
One can bound the *Joint Dependence Error*, or equivalently, the total correlation:

$$D_{\mathrm{KL}}\left(q(x^{\mathcal{S}}|x^{\bar{M}}), \prod_{l\in\mathcal{S}} q(x^l|x^{\bar{M}})\right) \leq \sum_{l\in\mathcal{S}} H(q(x^l|x^{\mathcal{S}})) - \max_{l\in\mathcal{S}} H(q(x^l|x^{\mathcal{S}})). \qquad (1)$$

When we assume low model error, this bound can be approximated with $p^\theta(x^l|x^{\mathcal{S}})$.

EB-Sampler

The *EB-Sampler* is a training-free, drop-in replacement solver offering unprecedented sampling efficiency at no cost, achieving $2-3x$ speed up on math and coding benchmarks.


1
2
3
4
5
$p^\theta(x^l|x^{\bar{M}})$
for
strings
Order of unmasking
Sort by error proxy
5
3
2
= err(
= err(
= err(
Choose how many to unmask
Bounded entropy
5
3
4
cumsum_entropy − cummax_entropy ≤ γ

EB-Sampler Step. Given and MDM with parameters θ, a state $x^{\bar{M}}$, an error proxy err and joint dependence threshold γ:

1. Compute the marginal conditionals $p^\theta(x^l|x^{\bar{M}})$
2. Compute $\mathrm{err}(p^\theta(x^l|x^{\bar{M}}))$ and sort tokens in ascending order
3. Unmask the maximal number of tokens such that Equation 1 is less than or equal γ

Experiments

Code and math benchmarks

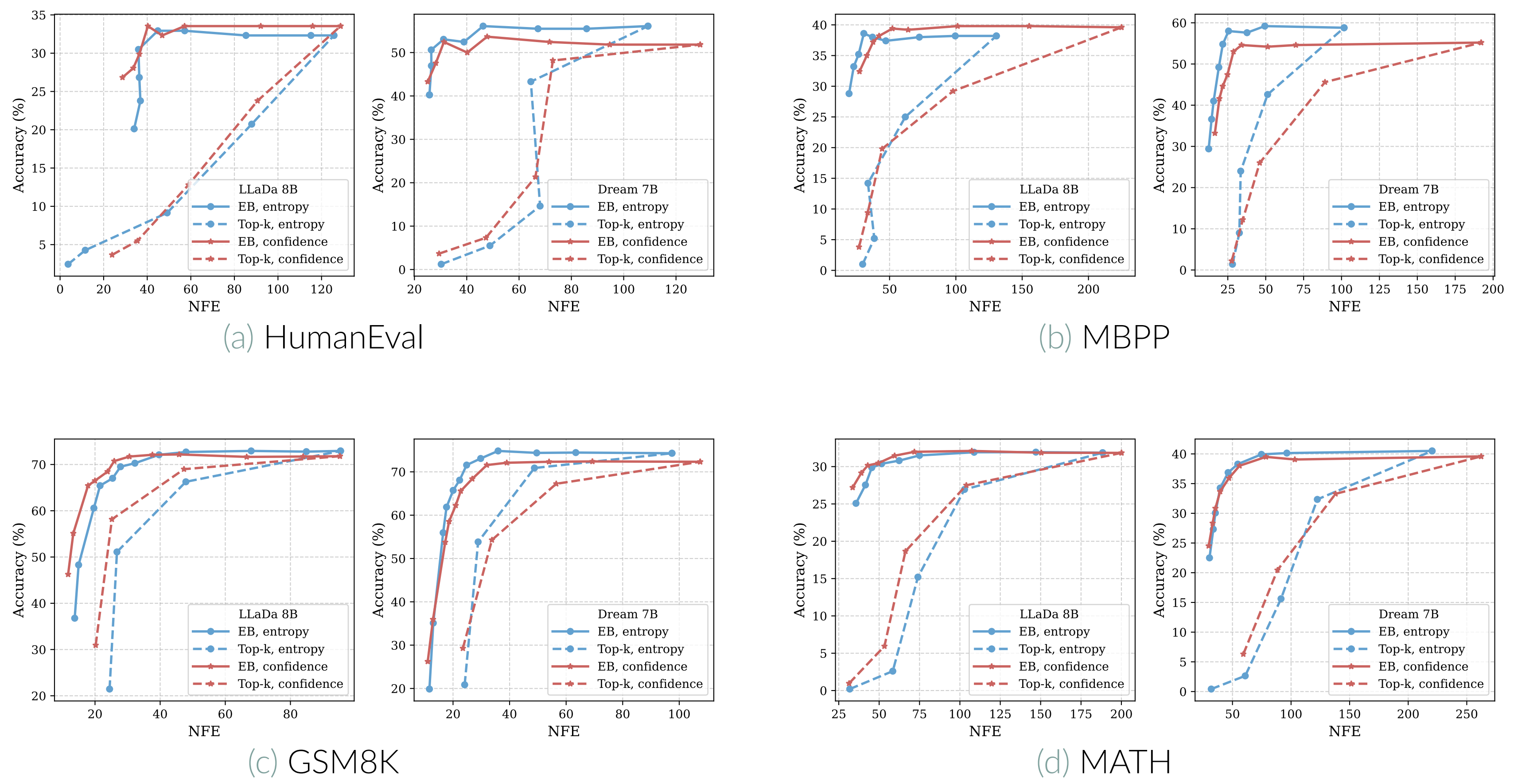


Figure 3. pass@1 accuracy vs. NFE with `generate_until` logic on code and math reasoning tasks.

Measuring efficiency gains of MDM samplers

We show accuracy against the effective generation speed of the model quantified via:

$$\text{Effective Tokens/Step} = \frac{\texttt{mean\_answer\_len}}{\texttt{mean\_NFE\_to\_condition}},$$

`mean_answer_len` - average #tokens from left to right until the `generate_until` answer markers.

`mean_NFE_to_condition` - the NFE required by the model to generate the answer until both `generate_until` answer markers appear and all tokens before the marker are unmasked.

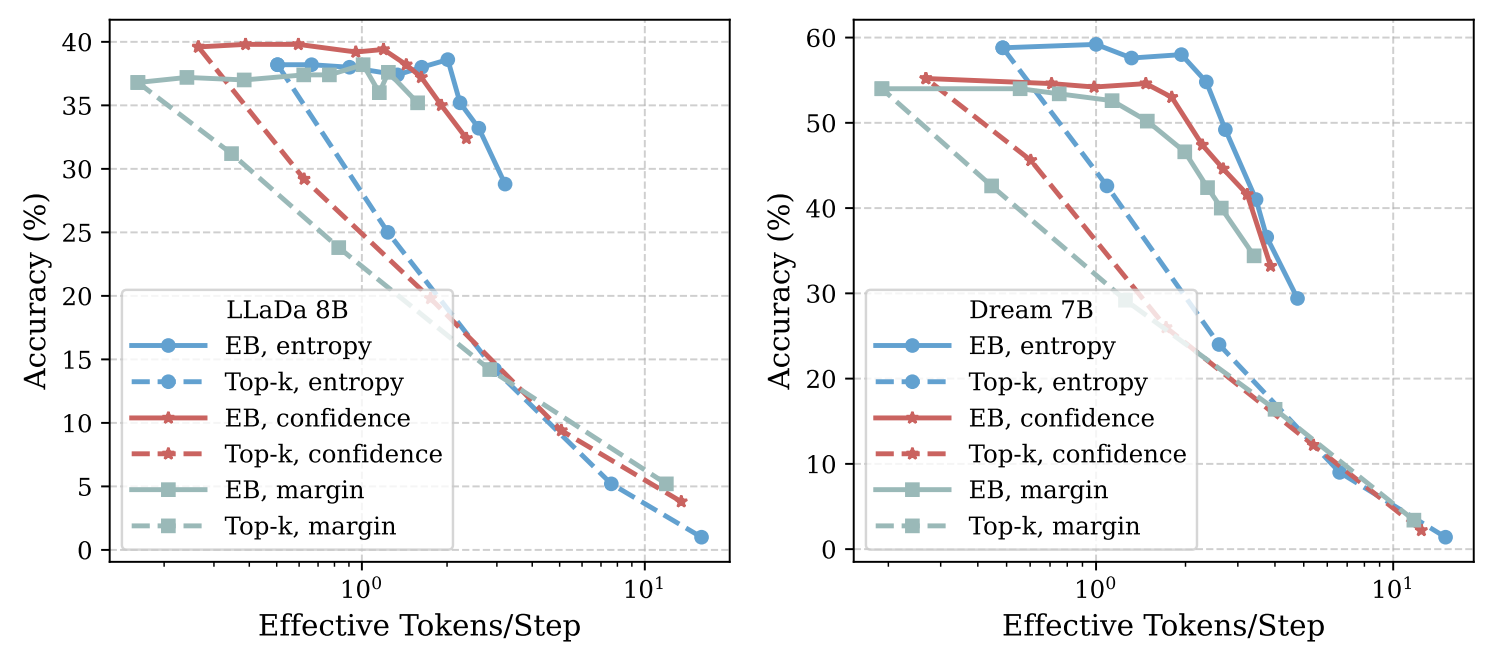


Figure 4. pass@1 accuracy vs. effective tokens/step on MBPP.

Surprisingly, in some cases, like the MBPP benchmark, the model finds it easier to unmask tokens that come after the answer, resulting in effective speed that is lower than 1 token per NFE.

Table 1. NFE and Speed-Ups for Dream 7B on MBPP for various evaluation schemes at roughly same best pass@1. For all configurations in the table the mean answer length is ~ 50 tokens.

	Full `max_gen_len` $= 512$ `generate_until` logic					`generate_until` logic + semi-AR (block_len=64)		
	pass@1	NFE	Speed-Up	NFE	Speed-Up	pass@1	NFE	Speed-Up
Top-1	58.8%	512	x 1	101.71	x 1	58.8%	64.59	x 1
EB, entropy, $\gamma = 0.001$	59.2%	174.57	x 2.93	49.39	x 2.05	59%	38.90	x 1.66
EB, entropy, $\gamma = 0.1$	58%	85.33	x 6.00	25.49	x 3.99	58.6%	21.19	x 3.05

References

[1] Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models, 2025.

[2] Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b, 2025.