EntRGi: Entropy-Aware Reward Guidance for Diffusion Language Models

Atula Tejaswi*, Litu Rout*, Constantine Caramanis, Sanjay Shakkottai, Sujay Sanghavi

Goal

- Given Discrete Diffusion LM (DLLM θ), Reward Model R
- **What is the best way to use gradient guidance to maximize R?**
- Challenges:
  - A. θ has a sampling step (unlike continuous diffusion)
  - B. R trained on discrete strings

Notations

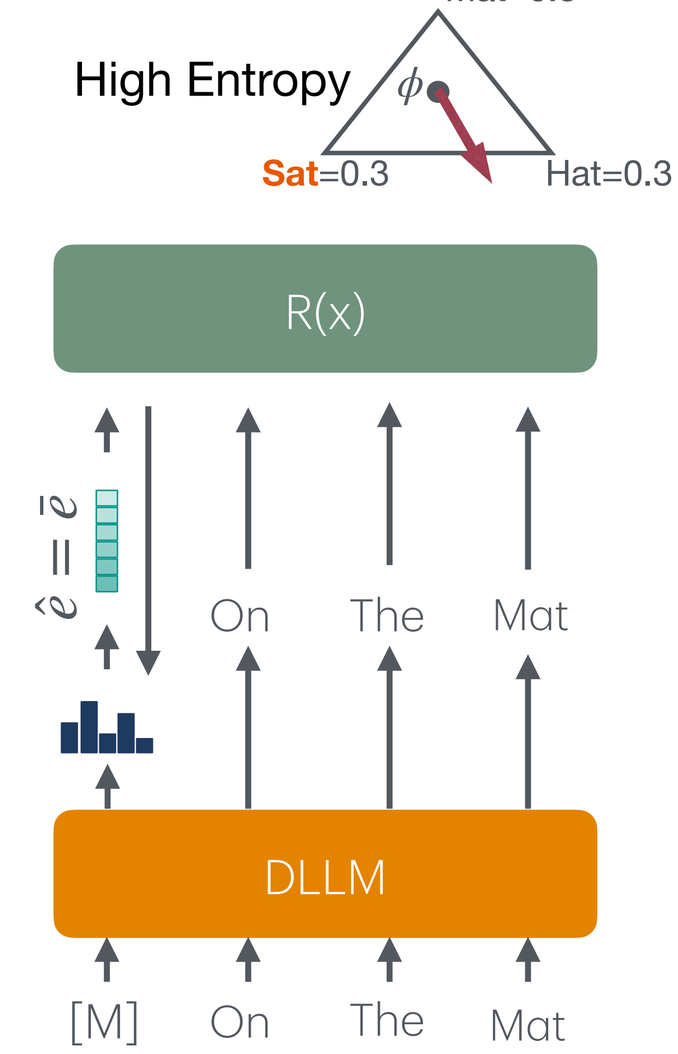
- E_R: embedding table of R
- Logits from DLLM: ϕ
- Vocabulary of DLLM/R: $\mathcal{V}$
- Sample Token Embedding: $\tilde{e} = E_R[x]$, $x \sim \text{softmax}(\phi)$ (eg: via Gumbel Noise)
- Reward Input Embedding: $\hat{e}$ (Method-Dependent)
- Expected Token Embedding: $\bar{e} = \sum_{i \in \mathcal{V}} \text{softmax}(\phi)_i \cdot E_R^i$
- Logit Update: $\phi \leftarrow \phi + \frac{1}{\beta} \nabla_\phi R(\hat{e})$

Expectation

Feeds the Expected embedding $\bar{e}$

Gradient Flows through $\bar{e}$

Inaccurate since R doesn't understand soft mixture $\bar{e}$



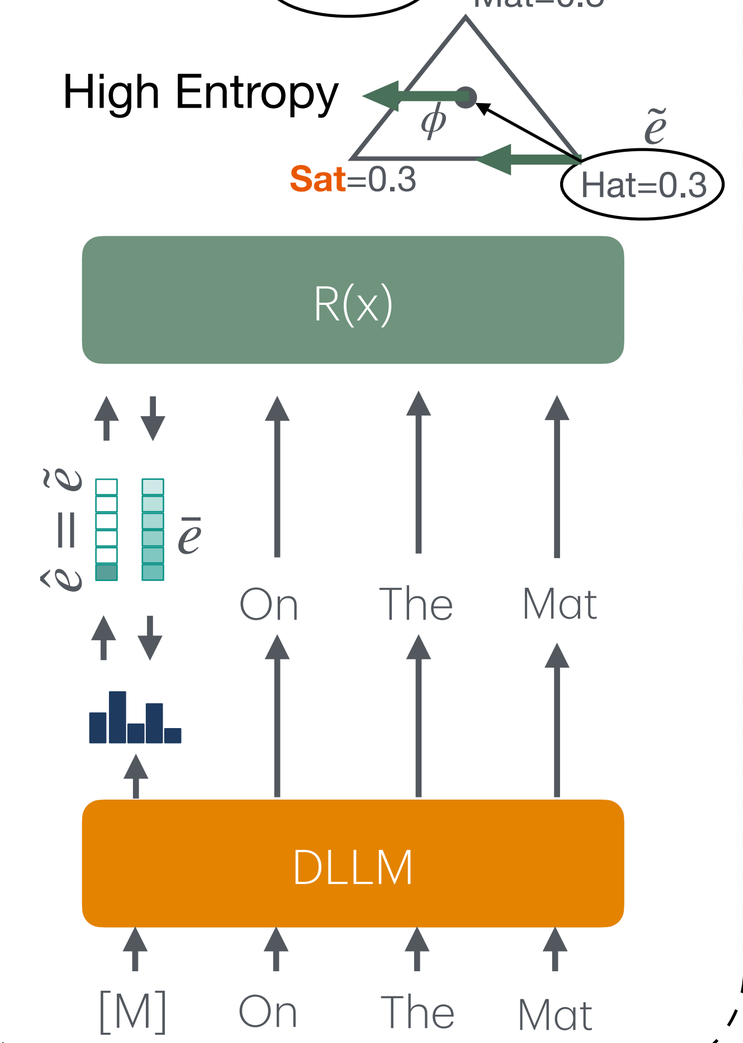
APS (Rout et al.)

Feeds the sampled embedding $\tilde{e}$

Gradient still flows through $\bar{e}$

$\hat{e} = \bar{e} + \text{stop-grad}(\tilde{e} - \bar{e})$

Inaccurate since R evaluated at $\tilde{e}$ but gradient propagated at $\bar{e}$



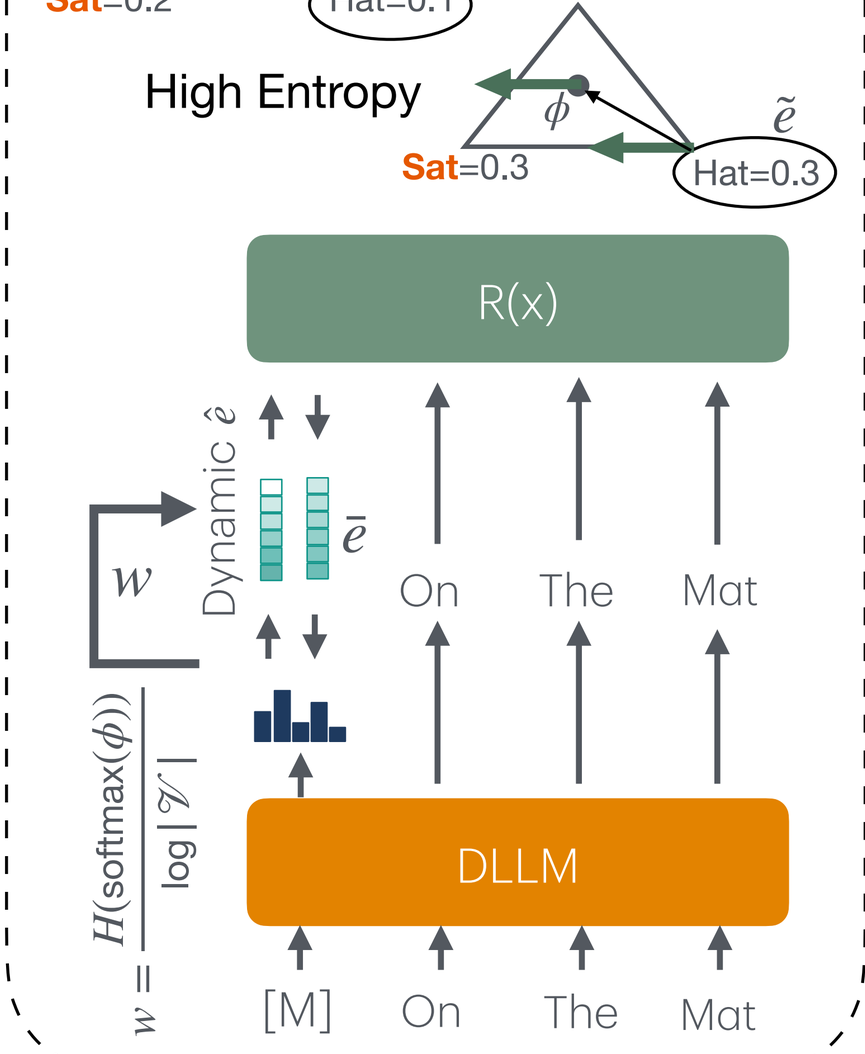
EntRGi (Ours)

Uses the observation that at low entropy, expected embedding is close to a real token

Feeds an entropy weighted mix to R

$\hat{e} = \bar{e} + \text{stop-grad}(w(\tilde{e} - \bar{e}))$

Gradient still flows through $\bar{e}$



Method	Reward-Bench-2			JudgeBench			RM-Bench		
	Top@1	**Avg@4**	**LMUnit**	**Top@1**	**Avg@4**	**LMUnit**	**Top@1**	**Avg@4**	**LMUnit**
Temperature ($\tau = 0.1$)									
BoN	0.18±0.22	0.05±0.23	3.74±0.04	0.00±0.15	-0.07±0.16	3.75±0.03	3.05±0.05	3.02±0.05	3.93±0.01
Expectation	2.19±0.19	**1.62**±0.17	4.12±0.03	0.68±0.19	**-0.06**±0.21	3.81±0.02	3.33±0.20	2.59±0.12	3.89±0.04
APS	2.95±0.21	1.47±0.20	4.19±0.01	**1.67**±0.11	-0.17±0.14	3.89±0.03	4.72±0.13	2.46±0.17	4.01±0.03
EntRGi	**3.07**±0.22	**1.62**±0.18	**4.22**±0.02	**1.73**±0.14	-0.11±0.18	**3.94**±0.01	**4.90**±0.13	**2.75**±0.14	**4.06**±0.01
Temperature ($\tau = 0.7$)									
BoN	2.99±0.23	1.38±0.29	4.15±0.02	1.65±0.18	-0.84±0.16	3.91±0.02	5.11±0.20	2.98±0.15	4.02±0.03
Expectation	**3.95**±0.28	**2.23**±0.24	4.22±0.02	2.30±0.08	**0.13**±0.07	**3.97**±0.01	5.45±0.16	3.29±0.13	4.02±0.03
APS	3.62±0.27	1.80±0.24	4.22±0.02	1.87±0.14	-0.63±0.10	3.93±0.02	5.11±0.14	2.66±0.15	4.00±0.02
EntRGi	**3.91**±0.30	**2.20**±0.26	**4.25**±0.02	**2.44**±0.06	0.02±0.10	**3.98**±0.02	**5.70**±0.12	**3.41**±0.14	**4.04**±0.01

DLLM:
Dream-7B-V0-Instruct

Reward Model:
Skywork-Reward-V2-{0.6B, 1.7B, 4B}

Try it Out!

Paper Code