NEURAL INFORMATION PROCESSING SYSTEMS

Uncertainty Quantification for Large Language Diffusion Models

Artem Vazhentsev[1], Vladislav Smirnov[1], David Li[1], Maxim Panov[1], Timothy Baldwin[1,2], Artem Shelmanov[1]

[1]MBZUAI [2]The University of Melbourne

Mohamed bin Zayed University of Artificial Intelligence

THE UNIVERSITY OF MELBOURNE

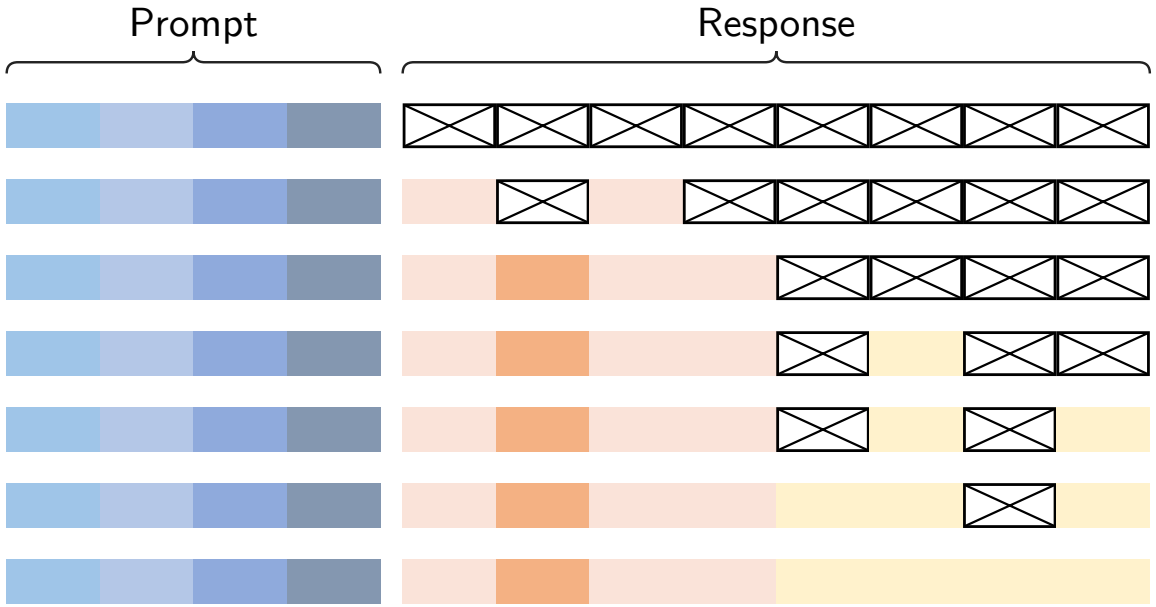
Generation in Large Language Diffusion Models

Iterative denoising

$$\underbrace{\mathbf{z}_T}_{\text{all [MASK]}} \rightarrow \cdots \rightarrow \underbrace{\mathbf{z}_0}_{\mathbf{y}}$$

$$\mathbf{z}_{t-1} \sim p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \mathbf{x})$$

- Generation starts from a fully masked sequence.
- Multiple tokens are predicted and revised in parallel.
- Intermediate states form a structured **denoising trajectory**.



Block Diffusion LLaDA

Why Standard UQ Falls Short for LLDMs

1. No autoregressive likelihood

$$p(\mathbf{y} \mid \mathbf{x}) = \prod_{l} p(y_l \mid \mathbf{y}_{<l}, \mathbf{x})$$

LLDMs lack this factorization and do not provide an exact tractable likelihood.

2. Repeated sampling is expensive

$$\mathbf{y}^{(i)} \sim p(\mathbf{y} \mid \mathbf{x}), \qquad i = 1, \ldots, M$$

Multiple generations undermine the **inference efficiency** of LLDMs.

Our key idea

Use intermediate denoising states $\{\mathbf{z}_t\}_{t=1}^{T}$ as **diffusion-specific uncertainty signals**.

Diffusion-Specific Signals for UQ

Key idea: reuse signals produced during denoising instead of sampling multiple outputs.

Likelihood-based signals

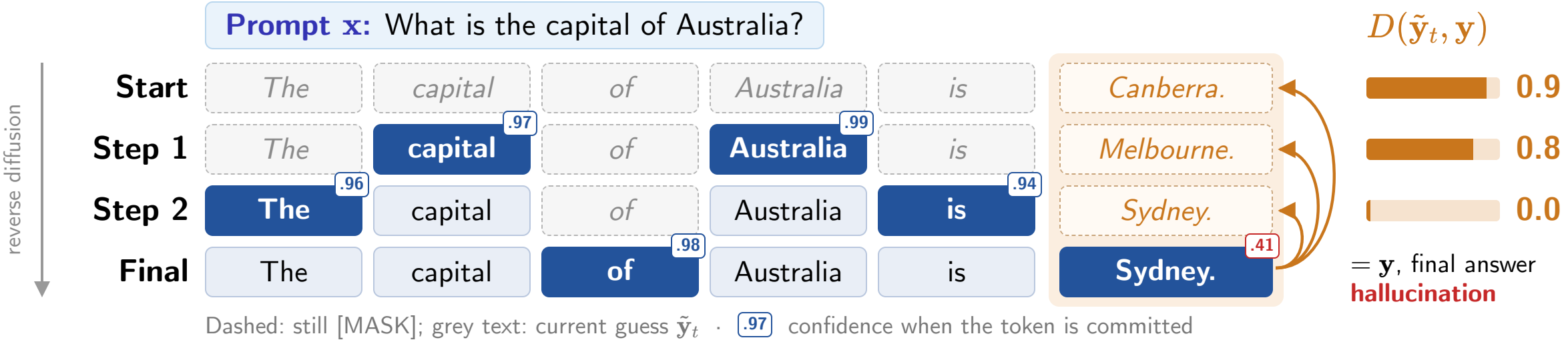
- **MCNLL:** masked diffusion likelihood estimate
- **Trajectory NLL / entropy:** uncertainty across denoising steps
- **Commitment NLL:** confidence when tokens are finalized

Trajectory-based signals

- **NFE:** generation complexity
- **Remasking:** frequency and instability of token predictions
- **Semantic dissimilarity:** semantic shifts during denoising trajectory

Denoising Trajectory Semantic Instability

Average trajectory dissimilarity to the final output: $\mathbf{u}_{\mathrm{AD}}(\mathbf{x}, \mathbf{y}) = \frac{1}{T} \sum_{t=1}^{T} D(\tilde{\mathbf{y}}_t, \mathbf{y})$



- High dissimilarity indicates substantial semantic revision and greater uncertainty.
- Unlike sampling-based UQ, it uses intermediate states from a **single generation**.

Theoretical grounding

Under calibration and an ideal semantic distance, expected trajectory dissimilarity lower-bounds the masked diffusion loss: $\mathbb{E}[\mathbf{u}_{\mathrm{AD}}(\mathbf{x}, \mathbf{y})] \leq \mathcal{L}(\theta)$.

D-CoCoA: Adapting CoCoA to LLDMs

CoCoA

$\mathbf{u}_{\text{inf}}$
sequence likelihood $p(\mathbf{y} \mid \mathbf{x})$

$\times$

$\mathbf{u}_{\text{cons}}$
consistency of M sampled answers

↓ no exact likelihood in LLDMs

↓ sampling is expensive

D-CoCoA

$\mathbf{u}_{\text{D-inf}}$
diffusion likelihood

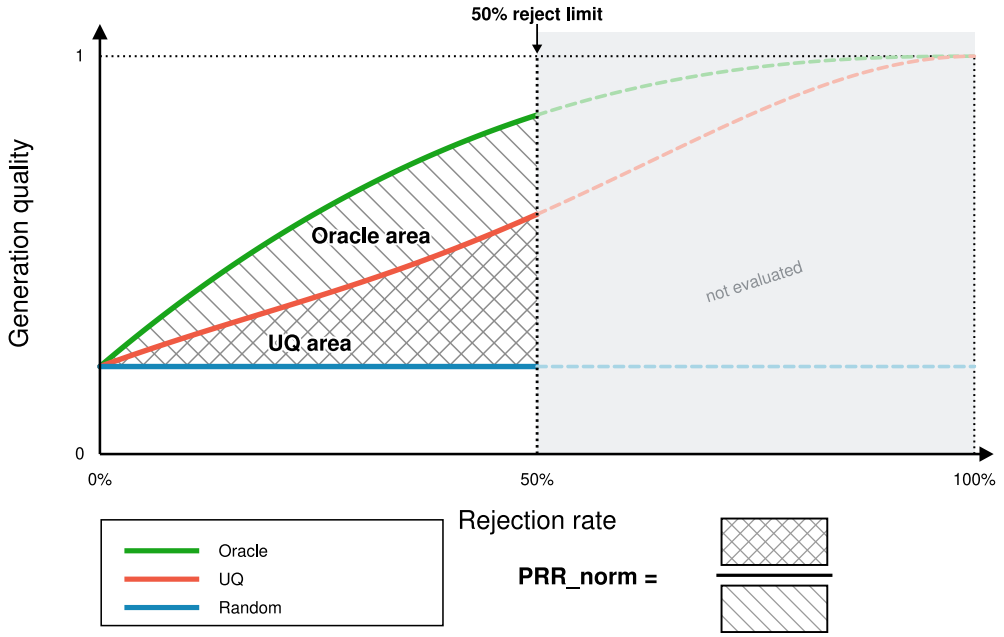
$\times$

$\mathbf{u}_{\text{AD}}$
trajectory dissimilarity, no sampling

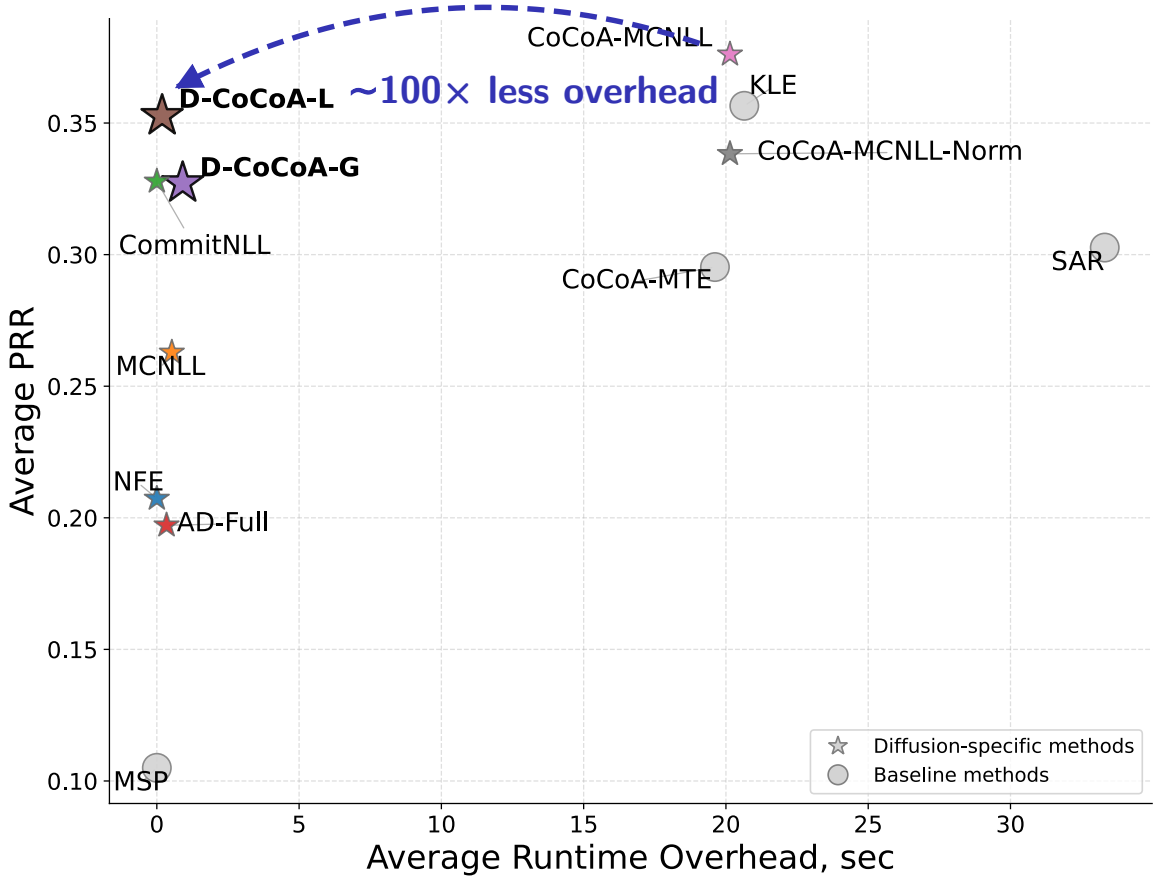
Variant	Diffusion likelihood	Trajectory term
D-CoCoA-L (local)	CommitNLL confidence when each token is committed	AD-Block dissimilarity within each block
D-CoCoA-G (global)	MCNLL masked diffusion likelihood, length-normalized	NFE · AD-Full denoising steps × whole-sequence dissimilarity

Experimental Setup

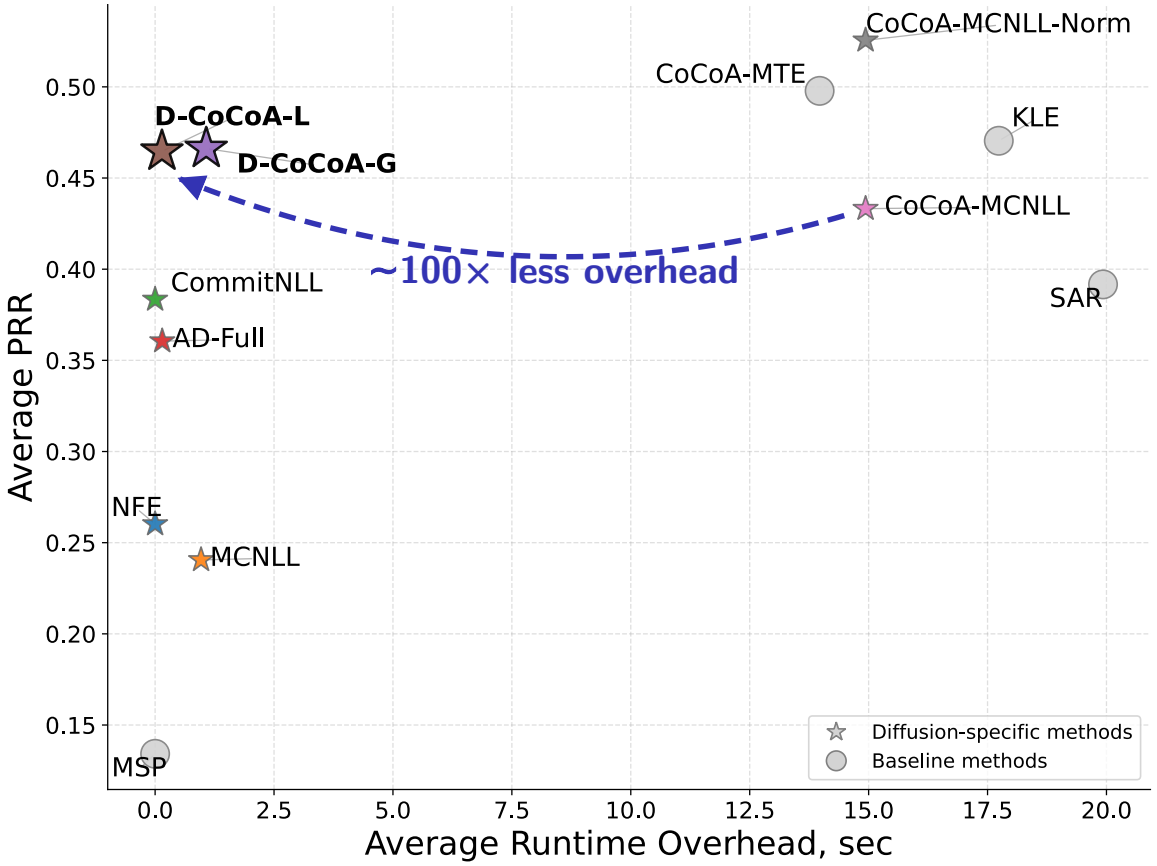
- **Task:** sequence-level selective generation
- **Datasets:**
  - QA with short free-form answers: CoQA, TriviaQA, MMLU
  - QA with long free-form answers: GSM8k
  - Translation: WMT14 Fr-En, WMT19 De-En
  - Summarization: XSum, SamSum
- **LLDMs:** LLaDA-1.5, Dream, ReFusion, DiffusionGemma
- **Metric:** PRR (50% max rejection), runtime in seconds



Results



(a) LLaDA-1.5



(b) Dream

- D-CoCoA methods approach the performance of sampling-based methods at ~100× lower runtime overhead

Takeaways

- Autoregressive logit scores perform only marginally better than random for LLDMs; use **diffusion-aware likelihood** (MCNLL).
- Intermediate denoising states can **replace sampling**.
- **D-CoCoA** approaches sampling-based methods at **∼100× lower overhead**.

mbzuai-nlp.github.io/uq-for-lldms

Fast inference and reliable UQ

LLDMs do not require a compromise between fast inference and reliable hallucination detection.