

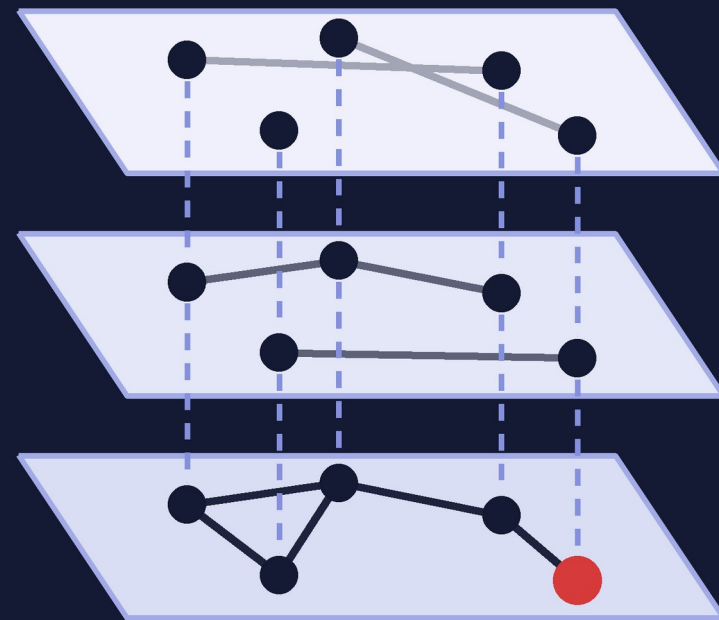
TDGNet: Hallucination Detection in Diffusion Language Models via Temporal Dynamic Graphs

Arshia Hemmat¹, Philip Torr¹, Yongqiang Chen^{2,3}, Junchi Yu¹

¹University of Oxford ²Carnegie Mellon University ³MBZUAI

EMNLP 2026 Oral

arxiv.org/abs/2602.08048 arshiahemmat.github.io/tdgnet

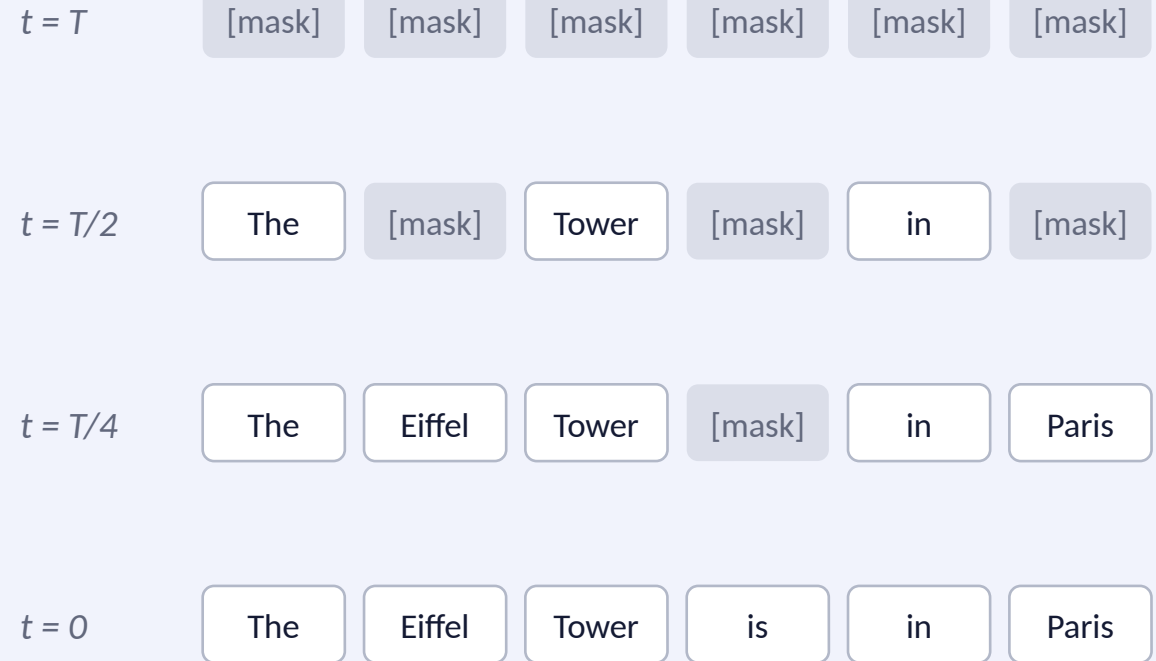


Diffusion LLMs write the whole answer at once

Start fully masked. All positions are denoised in parallel over T steps.

Bidirectional attention. Every step refines the global context of the whole sequence.

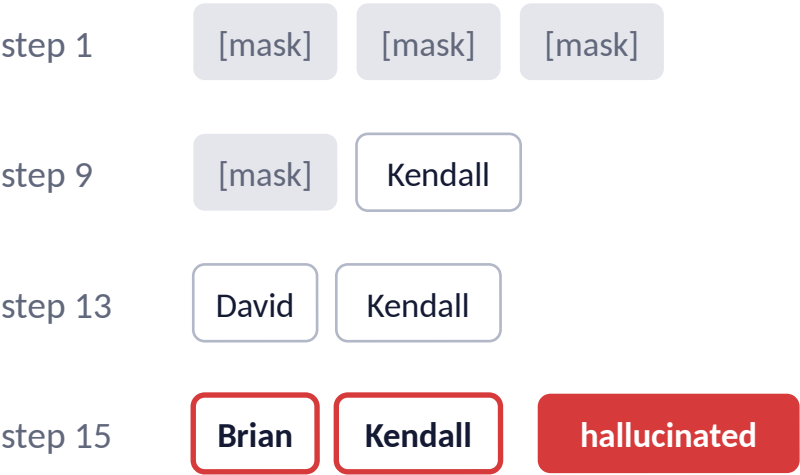
Already competitive. LLaDA-8B and Dream-7B match autoregressive models of similar size on many tasks.



Each step fills in more tokens; earlier choices can still change.

Hallucinations unfold over the denoising trajectory

“Who was manager of Sheffield Wednesday when they won the League Cup in 1991?”



Correct answer: Ron Atkinson

Self-correction
early noise resolves into the right answer

early

final

Correctness decay
a correct answer degrades into a wrong one

early

final

Semantic drift
a plausible guess drifts away from the truth

early

final

Persistent error
an error locks in early and never leaves

early

final

Example trace from the paper’s case studies (TriviaQA). Patterns are illustrative and can co-occur in one trajectory.

Existing detectors miss the trajectory



Output-based

Semantic Entropy, Lexical Similarity, LN-Entropy, Perplexity

Need several generations, and diffusion LLMs have no cheap single-pass likelihood.



Latent-based

EigenScore, CCS, TSV

Probe hidden states at one point and discard how they evolved across steps.



TDGNet (ours)

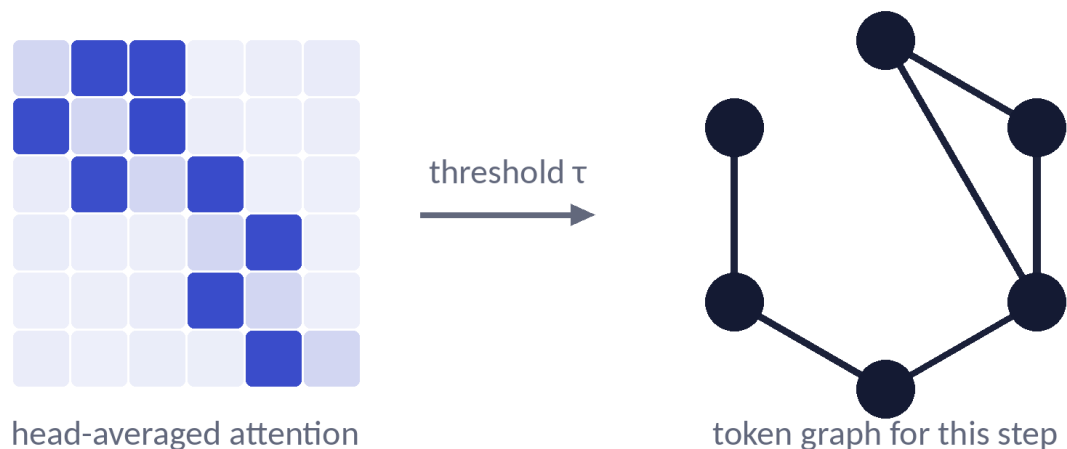
Temporal dynamic graphs over attention

Learns from the whole denoising trajectory, in a single generation.

Factuality evidence is spread across steps: it can appear, drift, or be corrected over time.



Key idea: read each denoising step as a token graph



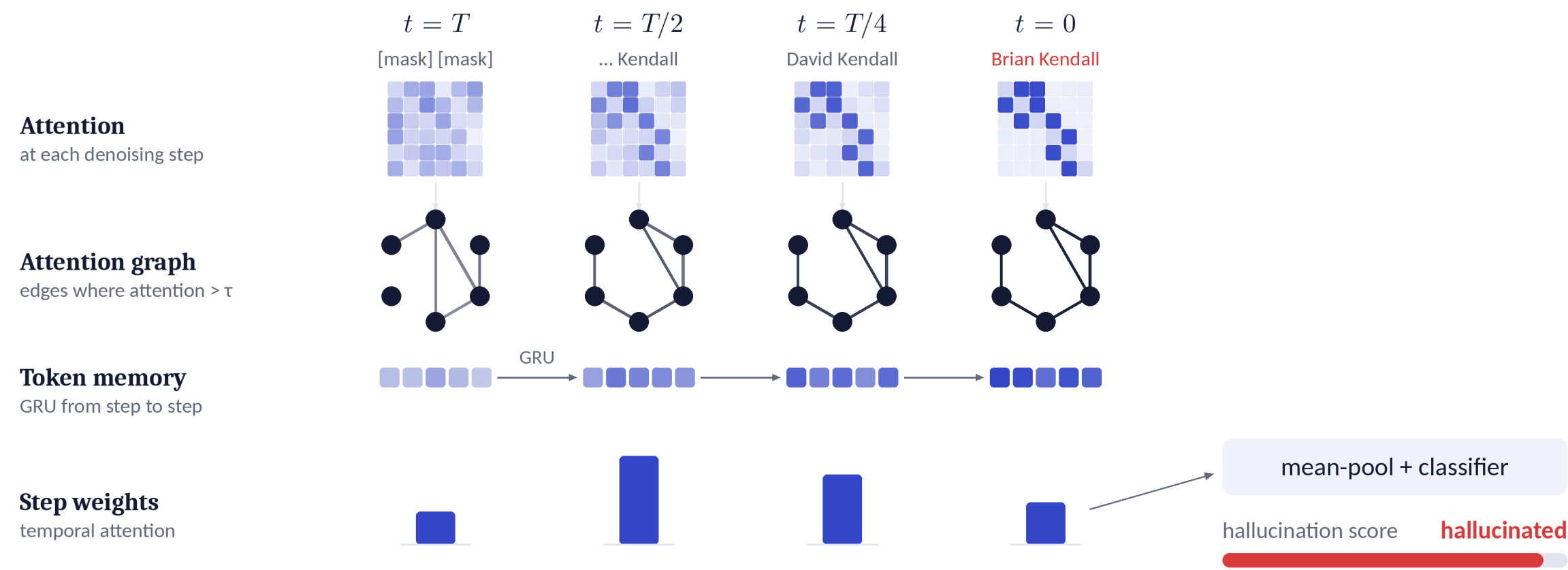
$$E^{(t)} = \{ (j, i) : \bar{A}_{ij}^{(t)} > \tau \}$$

An edge $j \rightarrow i$ when token i attends strongly to token j (head-averaged final-layer attention, $\tau = 0.05$).

Why attention?

- 1 Native.** The only per-step, per-token-pair signal the model exposes. No metric or extra model to choose.
- 2 Free.** Already computed in the forward pass. Building the graph is a single threshold.
- 3 On target.** Bidirectional attention is how diffusion LLMs refine context, so its evolution is what we want to track.

TDGNet tracks how token interactions evolve



Keyframes $t \in \{T, T/2, T/4, 0\}$. Node features: projected final-layer hidden states. Edge features: attention weights.

Three stages, trained end-to-end

1

Spatial aggregation

Message passing over each step's attention neighbourhood.

$$\bar{m}_i^{(t)} = \frac{1}{|\mathcal{N}^{(t)}(i)|} \sum_{j \in \mathcal{N}^{(t)}(i)} \psi(h_j^{(t)}, h_i^{(t)}, e_{ji}^{(t)})$$

2

Temporal memory

A GRU carries each token's evidence from step to step.

$$s_i^{(t)} = \text{GRU}(\bar{m}_i^{(t)}, s_i^{(t+1)})$$

3

Temporal attention readout

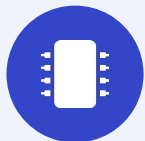
A learned query weighs the most informative steps per token.

$$z_i = \sum_t \alpha_i^{(t)} \text{Linear}(s_i^{(t)}), \quad \alpha_i^{(t)} \propto \exp(q^\top s_i^{(t)})$$

Response level: mean-pool the token embeddings, then a binary classifier (cross-entropy). Token level: classify each embedding directly.



Experimental setup



Models

LLaDA-8B-Instruct and Dream-7B-Instruct (main). LLaDA 1.5 (8B) and LLaDA 2.1-mini (16B MoE) for generalisation.



Datasets

Math, CommonsenseQA, HotpotQA, TriviaQA. Natural Questions and TriviaQA for token-level localisation.



Baselines

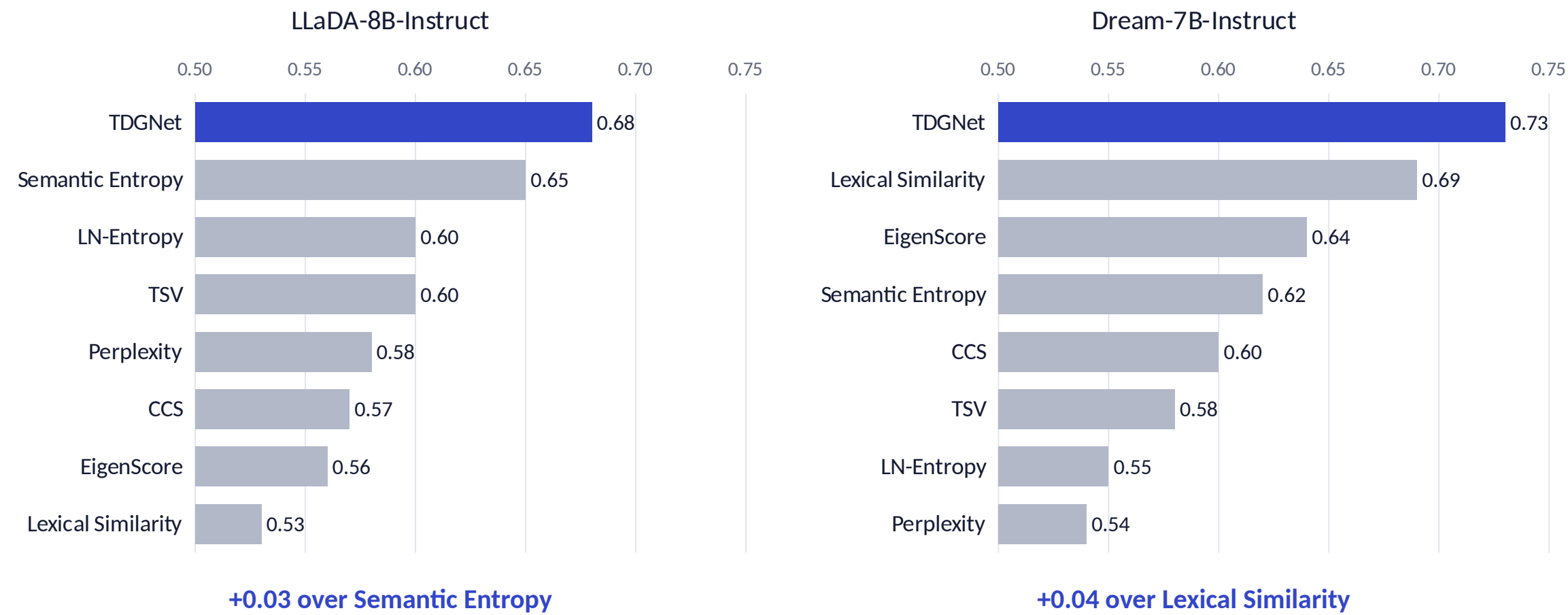
Output-based: Semantic Entropy, Lexical Similarity, LN-Entropy, Perplexity. Latent-based: EigenScore, TSV, CCS. Graph-based: CHARM.



Protocol

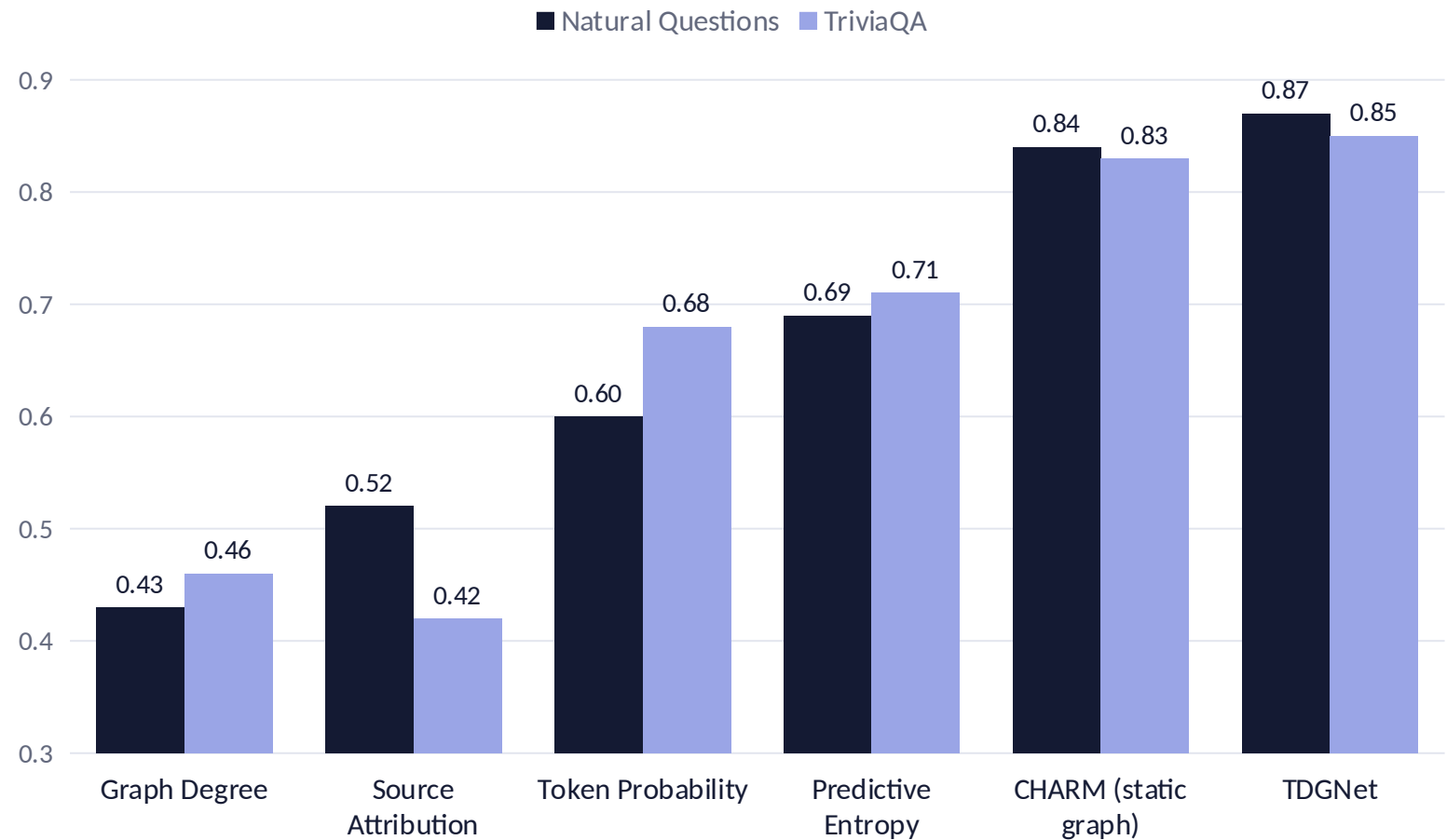
AUROC. 1,500 / 300 / 300 splits per dataset. 32 generated tokens, 16 denoising steps, $\tau = 0.05$.

TDGNet is best on average for both models



Average AUROC over Math, CSQA, HotpotQA and TriviaQA. Axes start at 0.5 (chance). Per-dataset results in the backup slides.

It also localises hallucinated tokens



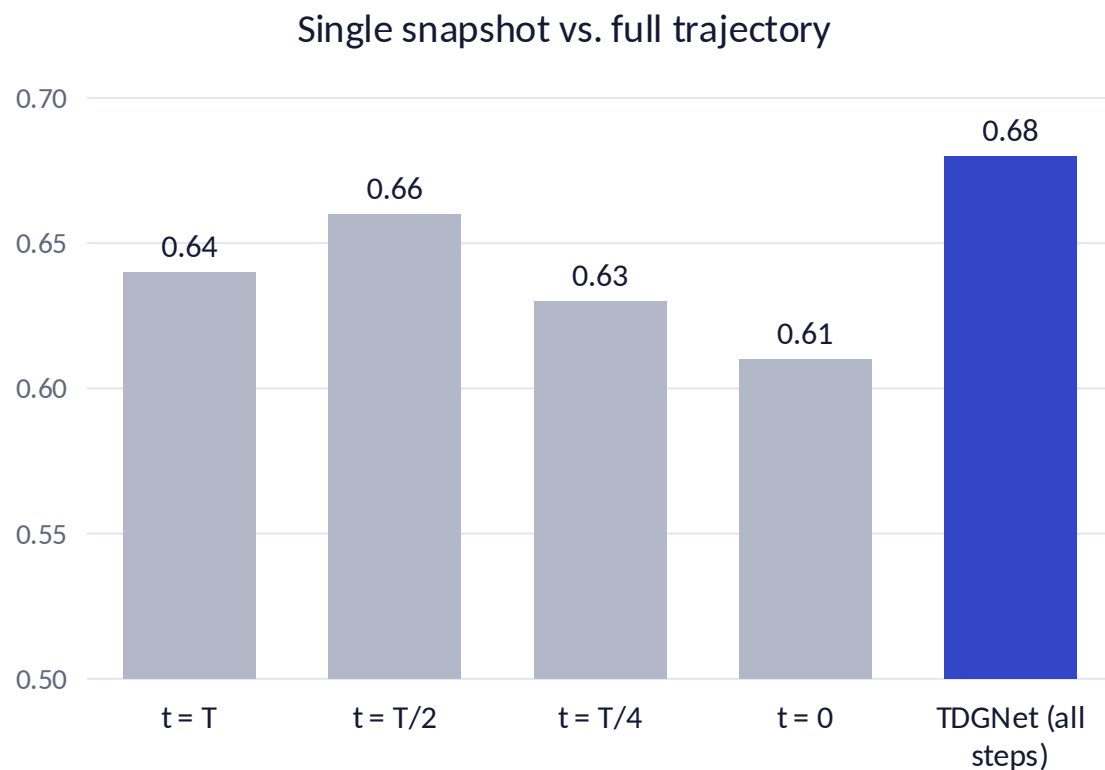
0.87

token-level AUROC on Natural Questions

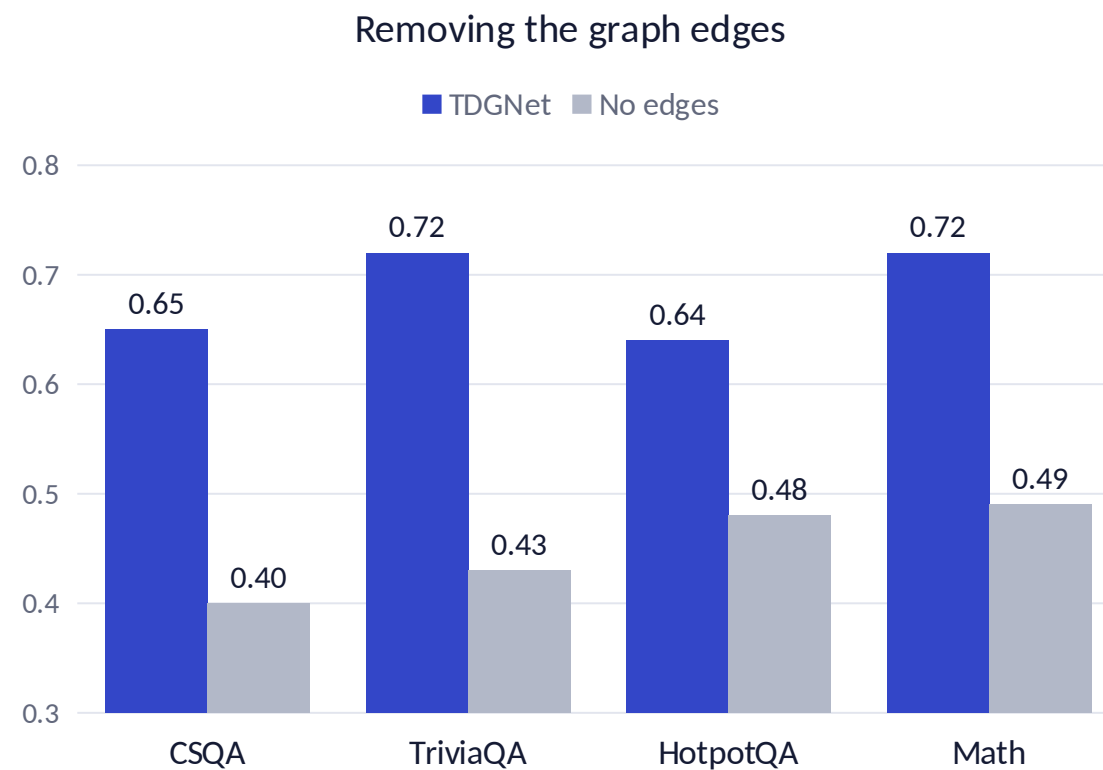
0.85

on TriviaQA, ahead of the static-graph CHARM (0.84 and 0.83)

Both time and structure matter



Static graphs at any single step trail the full trajectory.

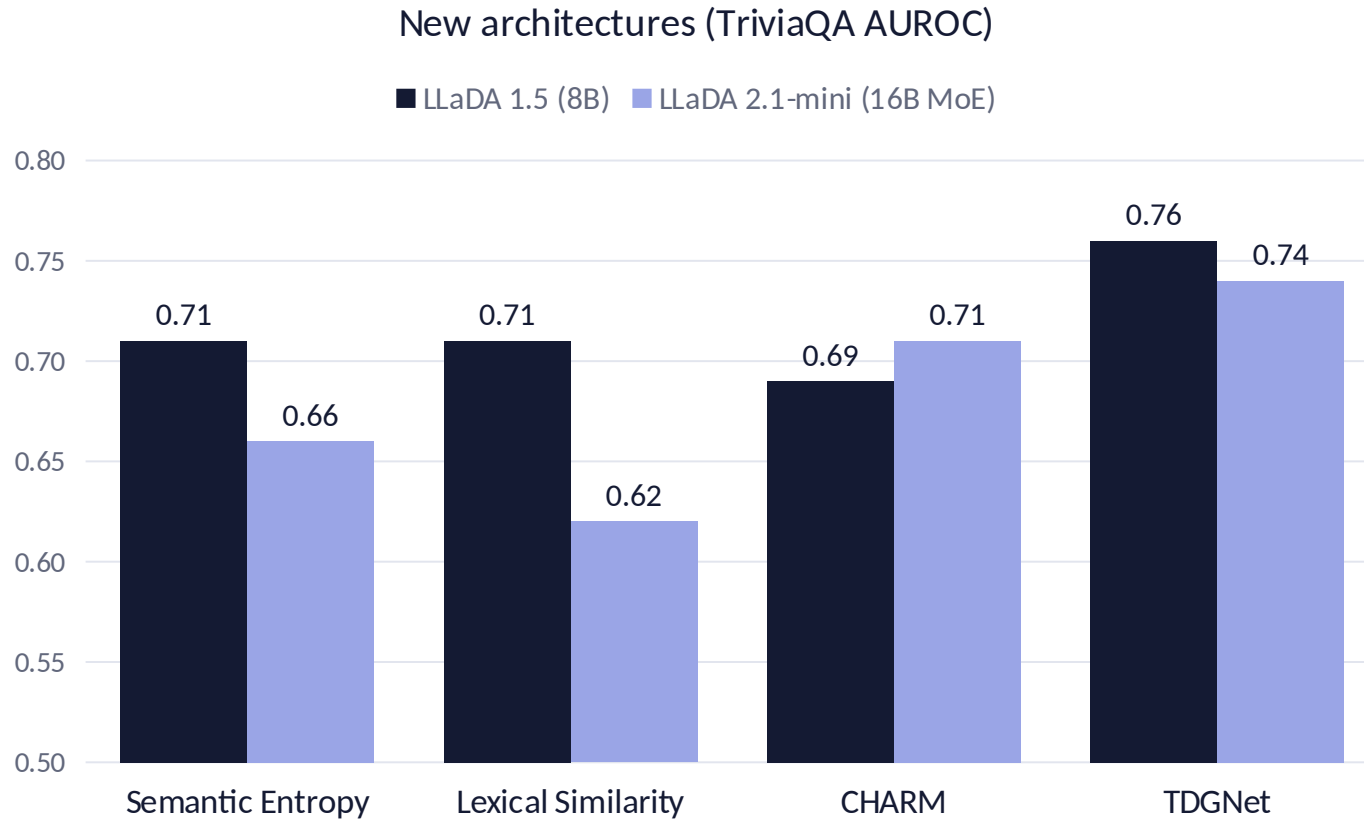


Without edges, AUROC falls to 0.40–0.49, near chance.

LLaDA-8B. Left: static-graph detector at one step (average AUROC). Right: TDGNet with the attention edges removed.



It generalises across models and datasets



Zero-shot across datasets

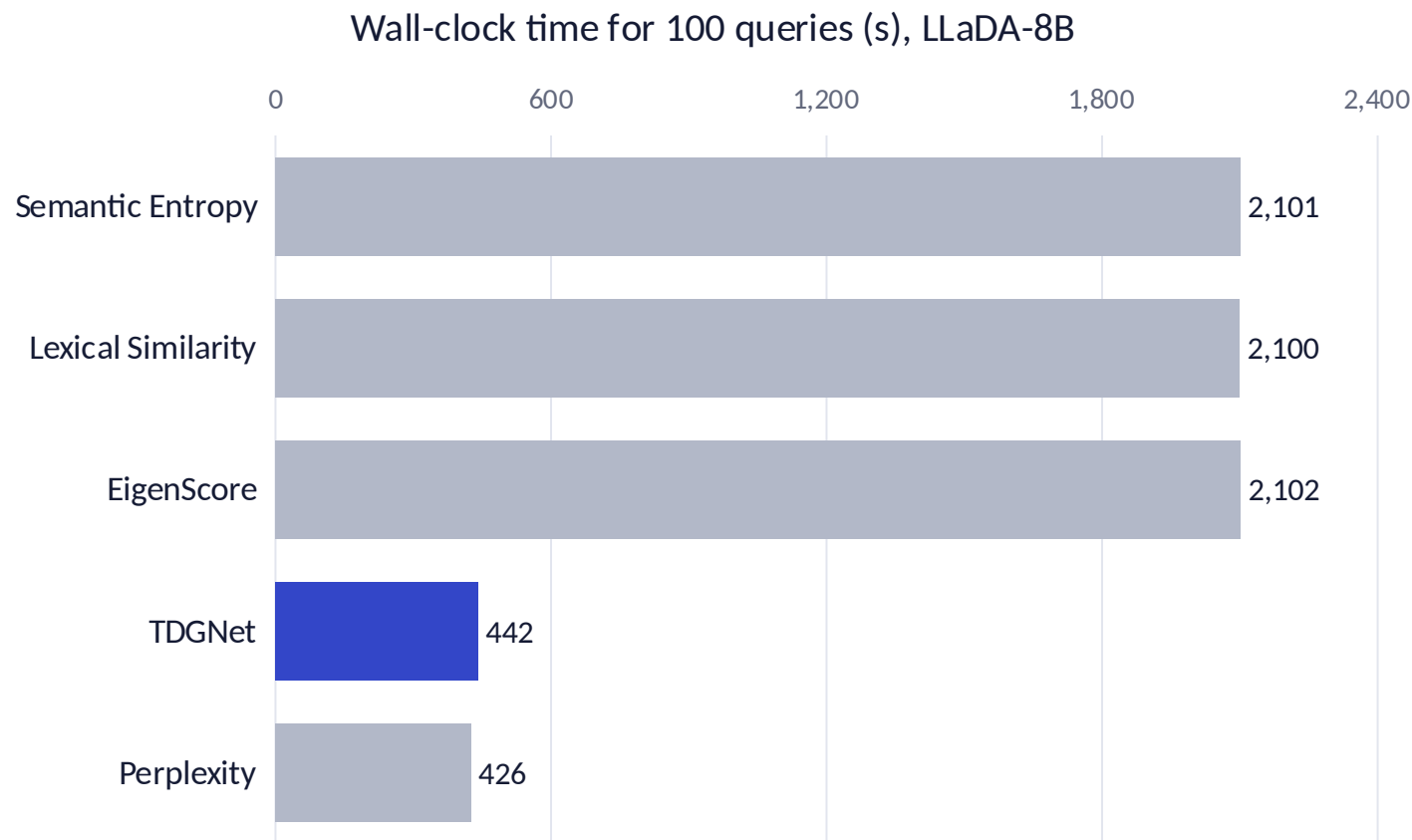
0.55–0.62

AUROC for TDGNet on all but one source-target pair, while CCS drops to 0.50–0.54 on distant pairs.

HotpotQA is the strongest source (0.59–0.62), suggesting multi-hop patterns transfer.



Near single-pass cost



4.8×

faster than Semantic Entropy



~160 ms

extra per query, about 4.4% of generation time



1

generation per query instead of 5



Takeaways



Factuality is a trajectory property

In diffusion LLMs, hallucinations appear, drift and self-correct across denoising steps.



Attention graphs over time capture it

Message passing, per-token memory and temporal attention, from a single generation.



Consistent gains

Best average AUROC on two models, token-level localisation, and transfer to new architectures.

Limitations

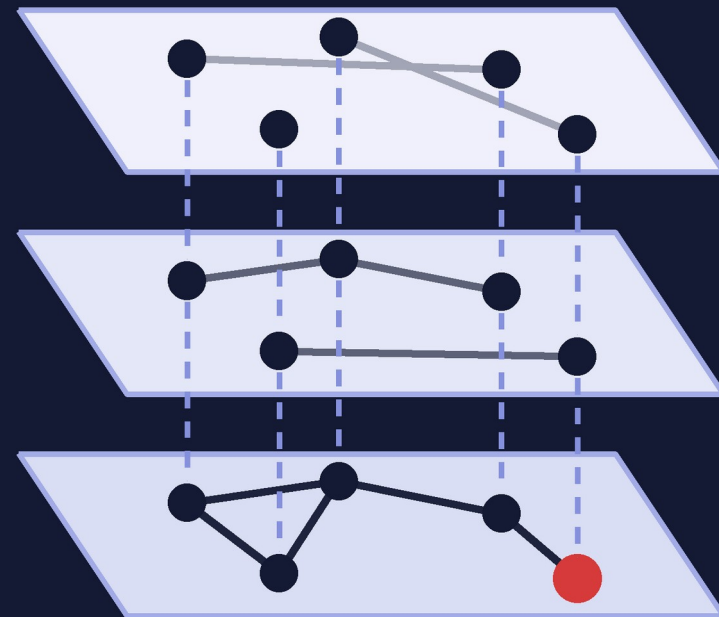
- QA-style benchmarks only: reliable factuality labels for long-form generation are scarce.
- Current open diffusion LLMs are unstable on long outputs, which blurs hallucination and fluency failures.

Thank you

Questions welcome

Paper arxiv.org/abs/2602.08048

Project arshiahemmat.github.io/tdgnet



Backup: per-dataset response-level AUROC

LLaDA-8B-Instruct

Method	Math	CSQA	Hotpot	Trivia	Avg
Semantic Entropy	0.68	0.64	0.61	0.66	0.65
Lexical Similarity	0.51	0.53	0.54	0.54	0.53
LN-Entropy	0.70	0.59	0.55	0.55	0.60
Perplexity	0.67	0.60	0.51	0.54	0.58
EigenScore	0.56	0.54	0.56	0.59	0.56
TSV	0.72	0.61	0.55	0.50	0.60
CCS	0.59	0.56	0.60	0.54	0.57
TDGNet	0.72	0.65	0.64	0.72	0.68

Dream-7B-Instruct

Method	Math	CSQA	Hotpot	Trivia	Avg
Semantic Entropy	0.59	0.50	0.68	0.69	0.62
Lexical Similarity	0.71	0.68	0.71	0.67	0.69
LN-Entropy	0.57	0.59	0.52	0.53	0.55
Perplexity	0.52	0.57	0.51	0.54	0.54
EigenScore	0.68	0.55	0.63	0.70	0.64
TSV	0.68	0.71	0.43	0.50	0.58
CCS	0.59	0.56	0.64	0.60	0.60
TDGNet	0.72	0.72	0.74	0.74	0.73

Table 1 of the paper. TDGNet is best or tied-best on every benchmark for both models.



Backup: comparison with concurrent TraceDet

Method	Math	CSQA	HotpotQA	TriviaQA	Avg
TraceDet (w/o masking) [†]	0.74	0.67	0.59	0.71	0.68
TDGNet ($\tau = 0.05$)	0.72	0.65	0.64	0.72	0.68
TDGNet ($\tau = 0.03$)	0.73	0.61	0.62	0.80	0.69

- Similar average AUROC. TDGNet is stronger on HotpotQA (+0.05) and TriviaQA (+0.09 at $\tau = 0.03$); TraceDet is stronger on Math and CSQA.
- TDGNet additionally supports token-level localisation, which entropy-trace methods do not.

[†]Reproduced from the specification in Chang et al. (2026); official code unreleased at the time of writing. LLaDA-8B.



Backup: sensitivity to the edge threshold τ

Threshold	Math	CSQA	HotpotQA	TriviaQA	Avg
$\tau = 0.03$	0.73	0.61	0.62	0.80	0.69
$\tau = 0.05$ (default)	0.72	0.65	0.64	0.72	0.68

- Robust for τ between 0.03 and 0.05 (average differs by 0.01). A lower τ keeps more edges and helps recall-oriented TriviaQA.
- At $\tau = 0.07$, over 97% of edges are removed and performance collapses to chance.

LLaDA-8B, response-level AUROC.

