NDAD: Negative-Direction Aware Decoding for Large Language Models via Controllable Hallucination Signal Injection

Panjia Qiu[1] *Mingyuan Fan*[1] *Cen Chen*[1†] *Daixin Wang*[2]

[1] East China Normal University [2] Ant Group

Github: https://github.com/SSSSSSilly/NDAD

Background

- **Large language models have achieved remarkable success in knowledge-intensive and reasoning tasks. However, they still frequently generate hallucinated content, i.e., fluent but factually incorrect or fabricated information, which limits their reliability in real-world applications.**

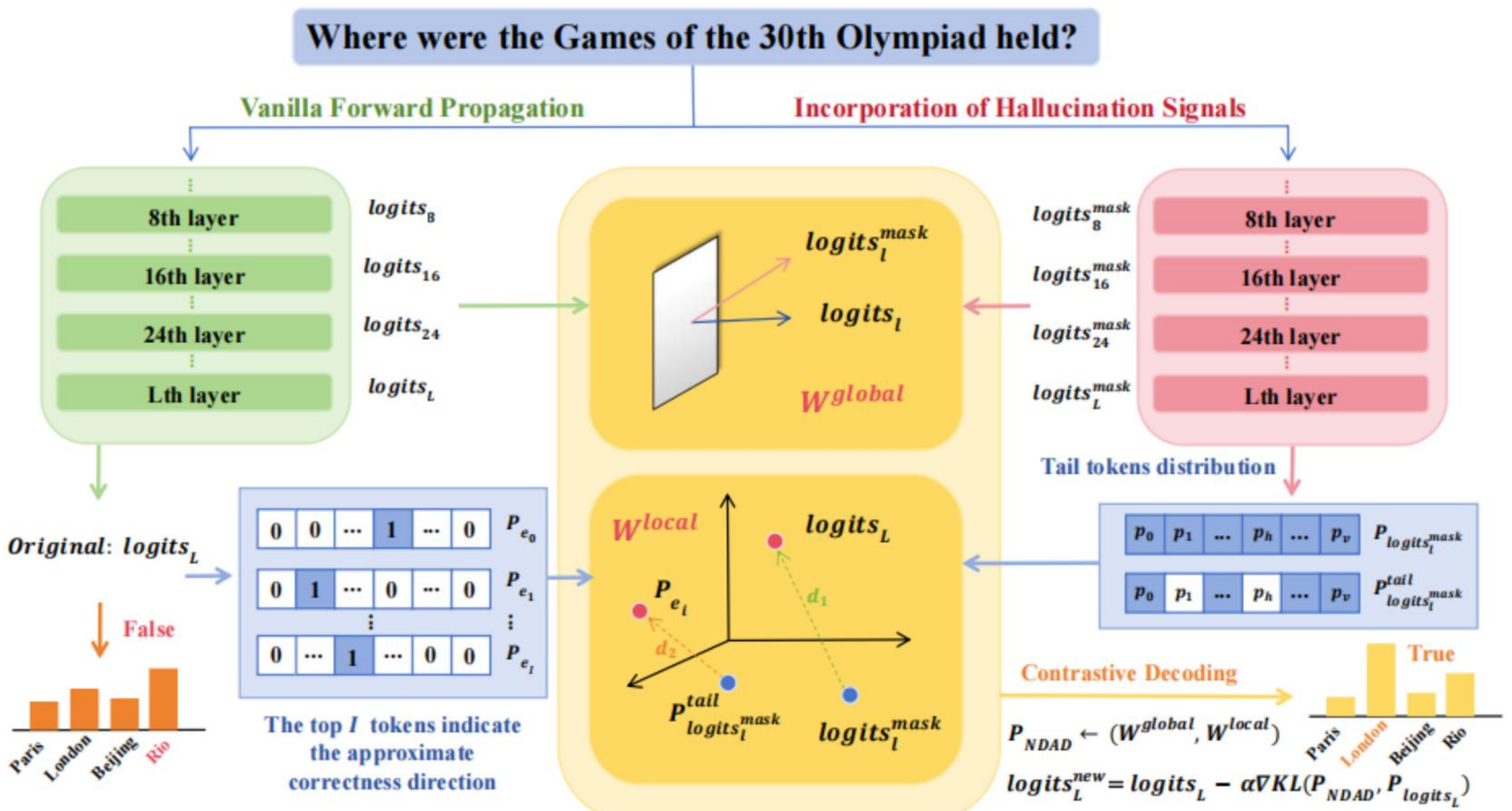
- **Existing approaches to mitigate hallucinations mainly fall into two categories: (1) Retrieval-based methods, which introduce external knowledge but increase latency and system complexity. (2) Training-based methods, which require additional supervision and may lack generalization.**

- **Despite these efforts, an important observation is that LLMs already encode rich factual knowledge in their internal representations.The key limitation lies in decoding, where such latent signals are not effectively utilized.**

Motivation:

Can we directly leverage internal signals during decoding to improve factuality, without retraining or external knowledge?

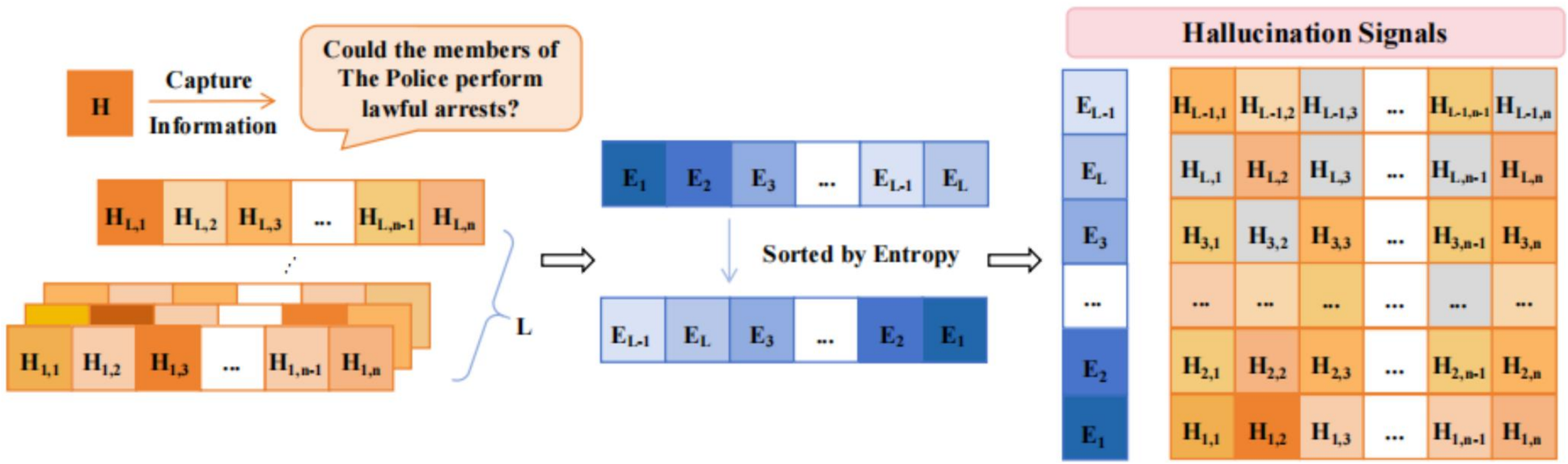
Method



NDAD augments vanilla decoding by injecting hallucination signals from masked attention heads.

It models both global consistency and local token evolution, and applies gradient-based adjustment to the final logits, steering generation away from hallucination-prone regions.

Step1-Hallucination Signals



1. Identify critical attention heads.
2. Rank layers via entropy-based importance.
3. Mask top heads to induce hallucination signals.
4. Construct hallucination-oriented distributions.

Step2-Dynamic Weighting Framework

Global Consistency:

- Measure the alignment between hallucination signals and original logits using cosine similarity.
- Higher consistency → more reliable signal.

Local Divergence:

Model the evolution of low-probability (tail) tokens:

$$W_{l,i}^{local} = max(cos((logits_L - logits_l^{mask}), (P_l^{tail} - P_{e_i})), 0)$$

- Focus on high-risk tokens
- Estimate their tendency to evolve into final outputs

Final Weight:

- Multiply global and local weights to highlight dominant signals and suppress noise.

Step3-Negative-Direction Aware Decoding

We incorporate the weighted hallucination signals to guide decoding away from hallucination-prone regions.

Latent Hallucination Distribution:

- **Aggregate weighted signals across layers to form a negative direction distribution.**

Gradient-based Logits Adjustment:

$$logits_L^{new} = logits_L - \alpha \Delta KL(P_{NDAD}, P_{logits_L})$$

- **Penalize divergence toward hallucination distribution.**
- **Suppress high-risk tokens while preserving confident predictions.**

Experiments

Method	TruthfluQA(MC)				Factor	CoT	
	MC1	MC2	MC3	Avg.	Wiki	StrQA	GSM8K
Llama2-7B-base	26.58	41.88	18.96	29.14	58.42	60.74	13.95
+DoLa-low	33.04	63.73	31.25	42.67	63.36	59.56	14.63
+DoLa-high	31.77	63.26	30.40	41.81	62.56	60.44	13.19
+AD	32.41	49.89	24.03	35.44	53.14	1.97	2.12
+SLED	34.15	62.57	31.89	42.87	67.00	61.27	14.63
+NDAD	**34.39**	**62.62**	**31.98**	**43.00**	**67.30**	**61.57**	**14.86**
Llama2-7B-chat	35.62	57.47	32.10	41.73	56.68	63.58	21.23
+DoLa-low	34.18	62.80	31.00	42.66	56.58	64.59	21.46
+DoLa-high	33.92	61.75	30.40	42.02	56.25	64.19	20.85
+AD	32.15	49.90	23.99	35.35	51.44	0.48	1.44
+SLED	**37.09**	**63.83**	**32.96**	**44.63**	64.80	64.50	21.53
+NDAD	36.84	63.42	32.93	44.40	**65.06**	**64.67**	**21.99**
Llama2-13B-base	27.59	43.14	19.53	30.09	63.79	65.98	28.81
+DoLa-low	31.57	62.48	30.41	41.49	65.70	66.46	28.51
+DoLa-high	29.38	63.92	33.62	42.31	52.84	60.83	11.90
+AD	32.15	49.90	23.99	35.35	58.18	2.01	0.00
+SLED	34.76	63.58	31.88	43.41	70.94	66.51	29.19
+NDAD	**34.88**	**63.60**	**31.97**	**43.48**	**71.18**	**66.81**	**29.26**
Llama2-13B-chat	36.47	63.06	32.77	44.10	61.96	69.65	36.69
+DoLa-low	34.27	63.27	31.36	42.97	60.69	69.48	35.48
+DoLa-high	31.82	62.55	31.13	41.83	54.81	66.51	33.21
+AD	32.15	49.90	23.99	35.35	56.71	23.14	0.00
+SLED	37.45	63.50	32.90	44.62	67.50	69.74	37.15
+NDAD	**37.58**	**63.63**	**33.02**	**44.74**	**67.74**	**69.96**	**37.30**

NDAD consistently improves factual accuracy across models and tasks, outperforming prior decoding methods such as DoLa and SLED.

The gains are especially evident in knowledge-intensive and reasoning benchmarks, demonstrating strong robustness and generalization

Contributions and Conclusions

Contributions:

- Propose NDAD, leveraging hallucination signals as negative directions for decoding.
- Design global + local dynamic weighting for controllable signal utilization.
- Apply gradient-based logits adjustment without retraining.
- Demonstrate consistent gains across models and tasks.

Conclusions:

- NDAD improves factuality by suppressing hallucination directions.
- Combines global consistency and local evolution for robustness.
- A lightweight, plug-and-play decoding method.