Hypergraph Modeling of Transformer Attention for Hallucination Detection

Mark Junjie Li[1,2], Zhishun Liu[1], Wei Wang[1] ✉, Yuanshan Lu[1], Jingxi Yang[1], Xianyu Bao[3], Jun Li[3], Sunjie Huang[3]

[1] Shenzhen University, Shenzhen, China · [2] Intelligent Computing Center, Shenzhen University, Shenzhen, China · [3] Shenzhen Academy of Inspection and Quarantine, Shenzhen, China

NeurIPS 2026 · Poster

Motivation

- Lack interpretability and incur high computational cost, while struggling with subtle, fact-consistent hallucinations.
- Rely on handcrafted heuristics and pairwise modeling, failing to capture higher-order dependencies in long-context and multi-hop reasoning.
- Pairwise modeling fails to capture higher-order structural dependencies, limiting effectiveness in long-context and multi-hop reasoning scenarios.

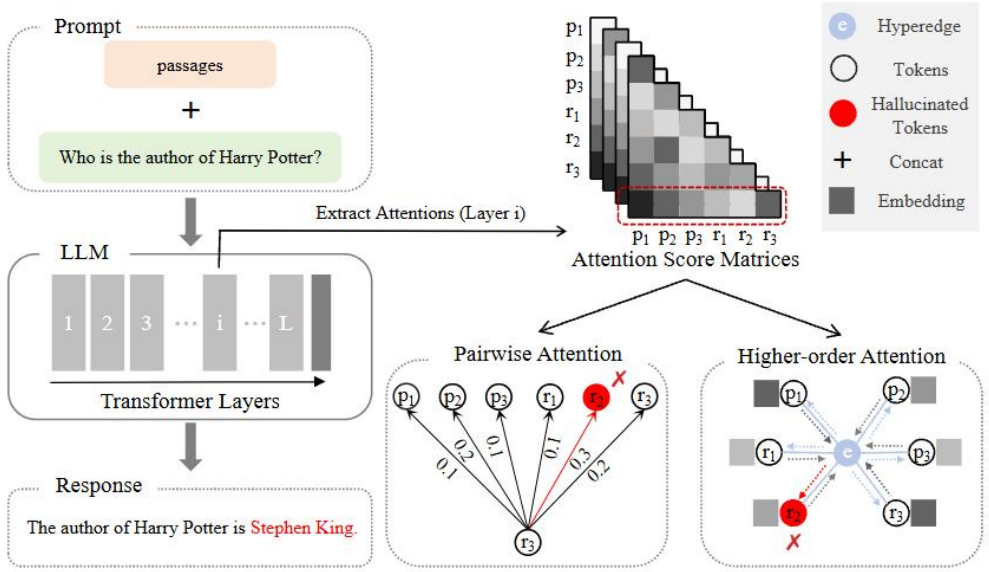


Figure 1: Pairwise vs. Higher-order Attention for Hallucination Detection.

Contributions

- Introduce attention-induced hypergraphs to capture higher-order, multi-token dependencies beyond pairwise relationships.
- Develop an HGNN-based framework for token-level hallucination detection through learnable message passing, eliminating handcrafted heuristics.
- Consistently outperform state-of-the-art detectors across multiple benchmarks, with robust zero-shot generalization to unseen tasks and settings.

Framework of AttnHyper

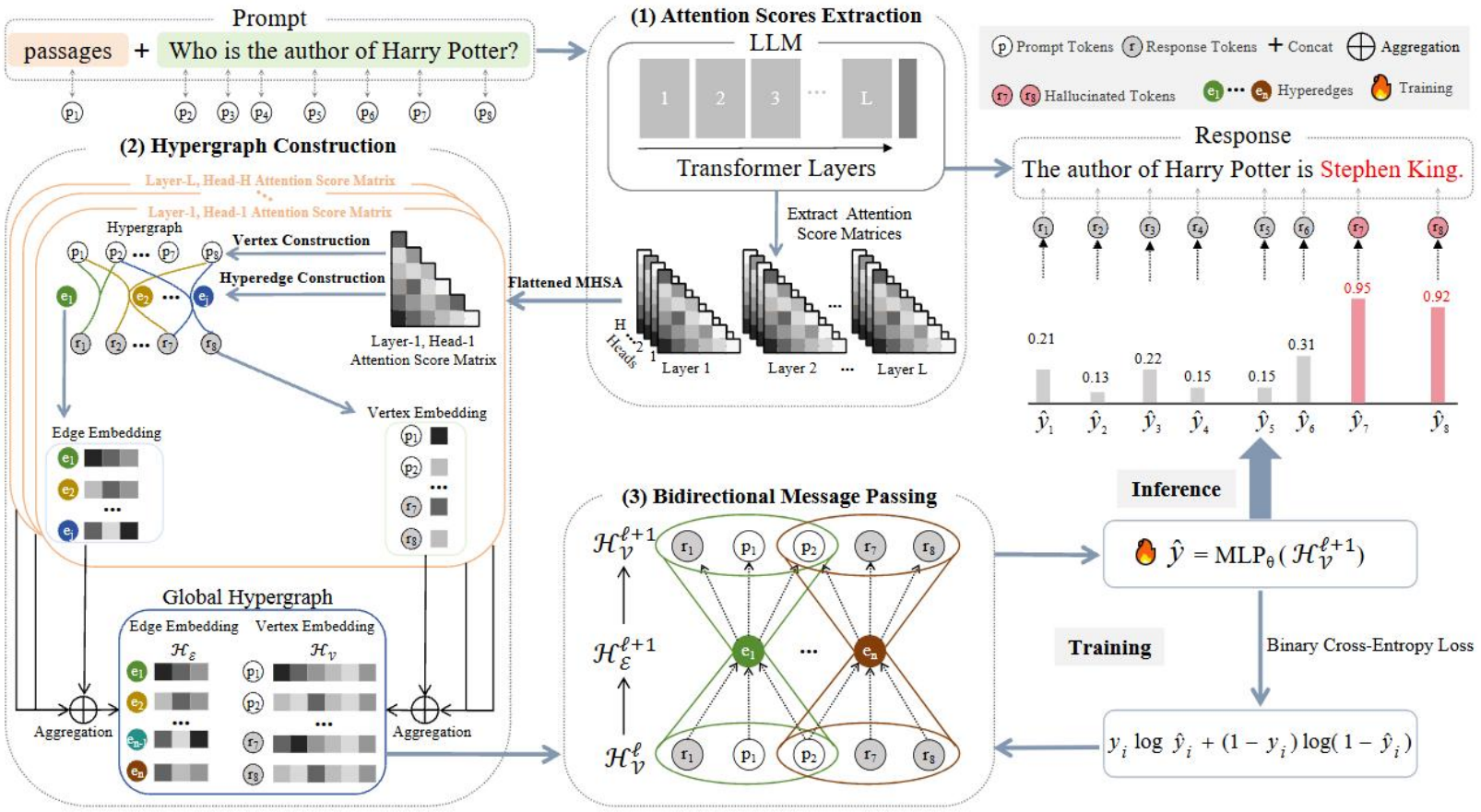
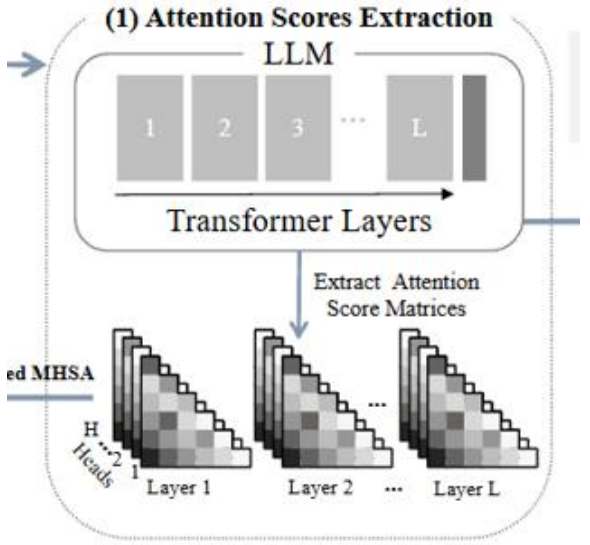


Figure 2: The overview framework of AttnHyper.

Stage 1: Attention Scores Extraction

Step1.Concatenate Prompt + Response tokens.

Step2.Extract attention scores across all layers and heads in one forward pass.



Stage 1:Attention Scores Extraction

Stage 2: Hypergraph Construction

Step1.Attention Flattening

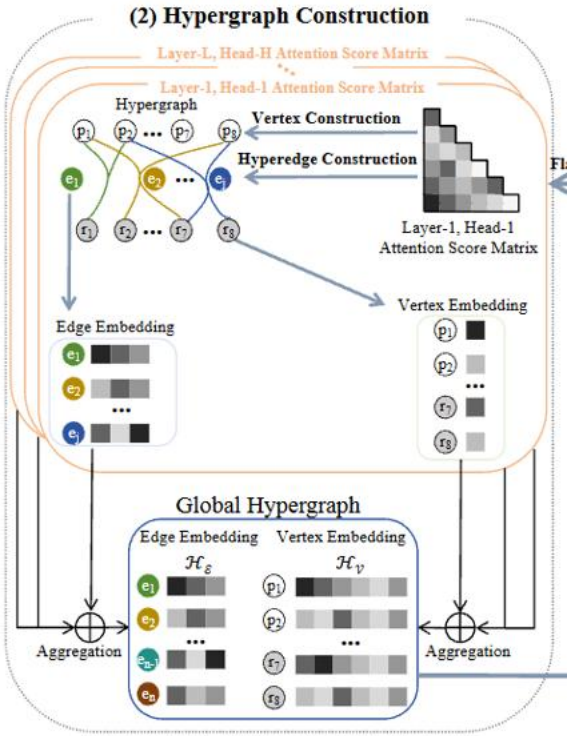
Transformer Layers × Attention Heads $\rightarrow \mathbf{A} \in \mathbf{R}^{(L \cdot H) \times (m+n) \times (m+n)}$ (5)

Step2.Vertex Construction

Prompt Tokens + Response Tokens → Vertices + Self-Attention Features

Step3.Hyperedge Construction

Response Token + Attended Preceding Tokens → Hyperedge Features: Mean / Max / Head ID



Stage 2:Hypergraph Construction

Stage 3: Bidirectional Message Passing

Step1.Node-to-Edge

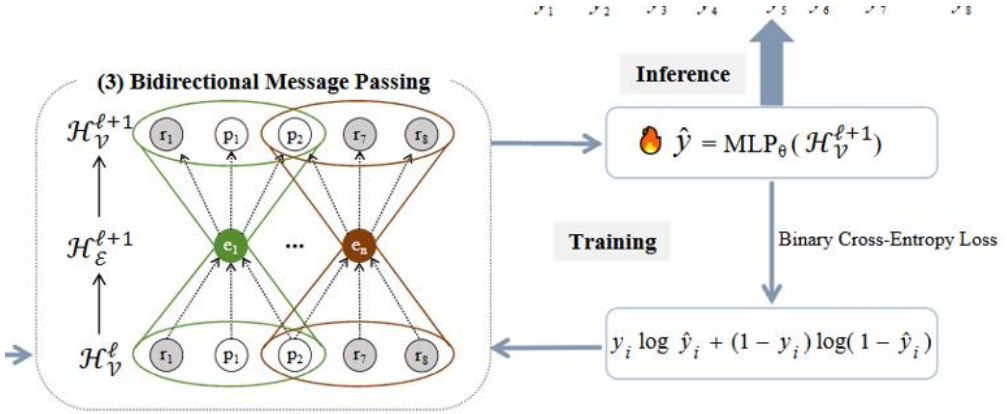
Aggregate connected node features to capture higher-order dependencies.

$$\mathrm{h}_{e}^{(l)} = \frac{1}{|e|} \sum_{v \in e} \mathbf{MLP}_{n \to e}\left(\mathbf{h}_{v}^{(l)}\right) \quad (16)$$

Step2.Edge-to-Node

Propagate hyperedge features and attributes back to nodes with residual connections and LayerNorm.

$$\mathrm{h}_{v}^{(l+1)} = \mathrm{h}_{v}^{l} + \mathbf{LN}\left(\mathbf{m}_{v}^{l}\right) \quad (17)$$



Stage 3:Bidirectional Message Passing

Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

深圳大学 SHENZHEN UNIVERSITY

Table 1: Performance comparison with state-of-the-art baselines on AUROC (%) for RAGTruth. Higher values indicate better performance. The best and second-best results are in bold and underlined, respectively. For our method, mean and standard deviation are computed across five random seeds.

Model	Type	Method	QA	Summary	Data2txt
Mistral-7B	Logits	Perplexity	59.14	64.66	33.91
	Hidden	ICR	60.96	55.66	53.33
		TSV	79.43	66.07	68.86
	Attention	LapEig	76.31	68.39	66.65
		Lookback	<u>91.35</u>	<u>77.75</u>	<u>89.96</u>
		AttnHyper	**96.18**$\pm$0.1	**87.89**$\pm$0.8	**94.04**$\pm$0.4
LLaMA-2-7B	Logits	Perplexity	54.29	60.81	36.85
	Hidden	ICR	58.86	53.86	60.18
		TSV	78.05	63.57	66.52
	Attention	LapEig	79.63	63.34	65.44
		Lookback	<u>85.09</u>	<u>69.65</u>	<u>85.05</u>
		AttnHyper	**90.18**$\pm$0.1	**71.54**$\pm$0.4	**89.51**$\pm$0.3

Experiments

Table 2: Cross-dataset transferability (AUROC, %) on the Mistral-7B-Instruct.

Train	Method	QA	Summary	Data2txt
QA	Lookback Lens	91.35	74.43	69.25
	ICR Probe	60.96	57.03	52.37
	AttnHyper	**96.18**	**78.41**	**77.53**
Summary	Lookback Lens	84.06	77.75	68.87
	ICR Probe	62.68	55.66	52.77
	AttnHyper	**94.41**	**87.89**	**88.94**
Data2txt	Lookback Lens	83.80	71.73	89.96
	ICR Probe	62.38	50.51	53.33
	AttnHyper	**92.24**	**84.15**	**94.04**

Experiments

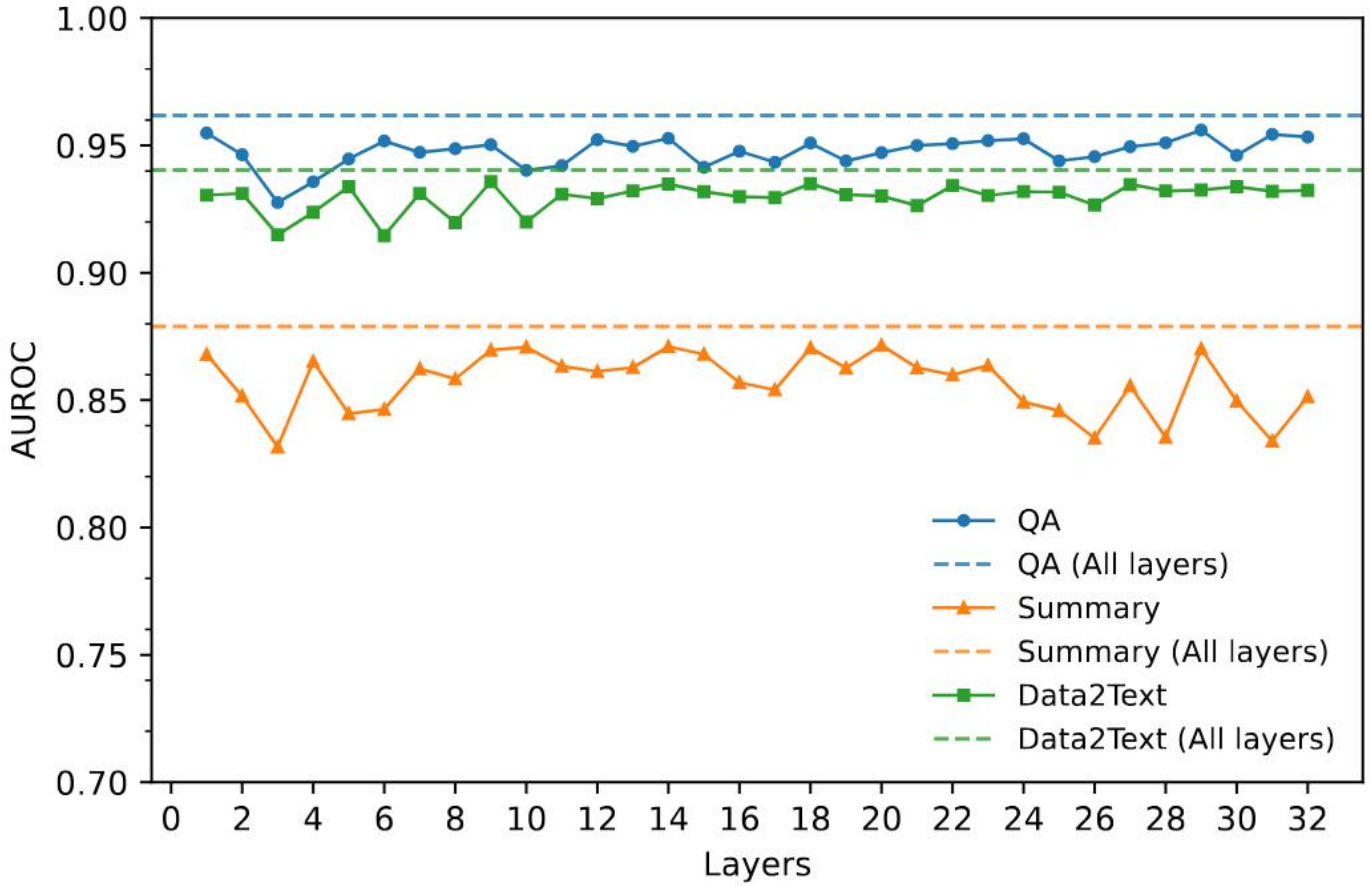


Figure 3: Average AUROC across transformer layers for different tasks (QA, Summary, and Data2txt) on Mistral-7B-Instruct.

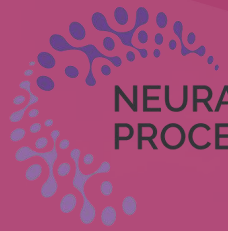
Conclusion

- Novel Framework: Introduce AttnHyper, the first framework to model Transformer attention patterns as hypergraphs for LLM hallucination detection.
- Higher-Order Dependencies: Capture complex token interactions by connecting multiple co-attended tokens through hyperedges.
- Bidirectional Information Aggregation: Leverage bidirectional message passing to integrate information across attention heads and layers.
- Superior Performance: Outperform existing methods on the RAGTruth benchmark with strong zero-shot generalization.
- New Interpretability Perspective: Reveal structural patterns in attention, bridging low-level token interactions and high-level reasoning.
- Broader Impact: Provide new insights into LLM internal mechanisms and pave the way for safer and more trustworthy language models.



深圳大学

SHENZHEN UNIVERSITY



NEURAL INFORMATION
PROCESSING SYSTEMS

Thanks !

