Reading Attribution from Attention: Evidence Heads as Latent Attribution Mechanisms in LLMs

Qi Wu[1], Jianfeng Qu[2*], Yanzhe Ji[3], Peng-Fei Zhang[4], Siyu Li[2], Wei Chen[1], Zhixu Li[2], Jiajie Xu[1]

[1]Soochow University · [2]Renmin University of China · [3]Australian National University · [4]Tongji University

NEURAL INFORMATION PROCESSING SYSTEMS

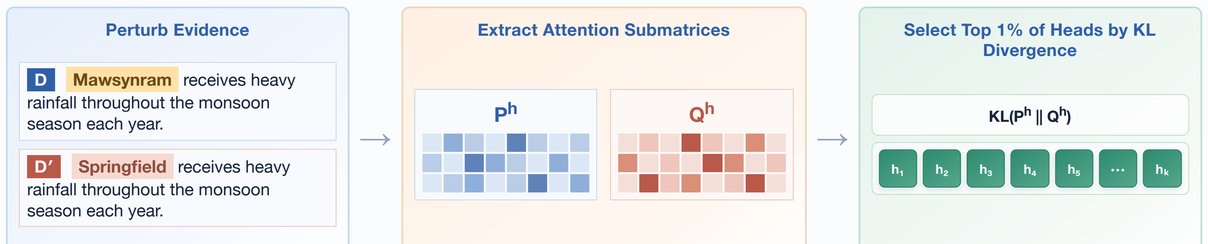
Scan for the full paper

1 Motivation

- **Reliable evidence attribution is essential for trustworthy LLMs.** In multi-document question answering, models must identify which sources support each answer statement, especially in applications such as retrieval-augmented generation and factual verification.
- **Current attribution methods have important limitations.** Prompted citations can be missing, inconsistent, or unsupported, while training-based approaches require annotated data and additional training and provide limited insight into how evidence is selected.
- **LLMs may already encode useful attribution signals internally.** Since attention heads can specialize in particular functions, identifying heads that respond to evidence could enable a more interpretable, training-free approach to attribution.

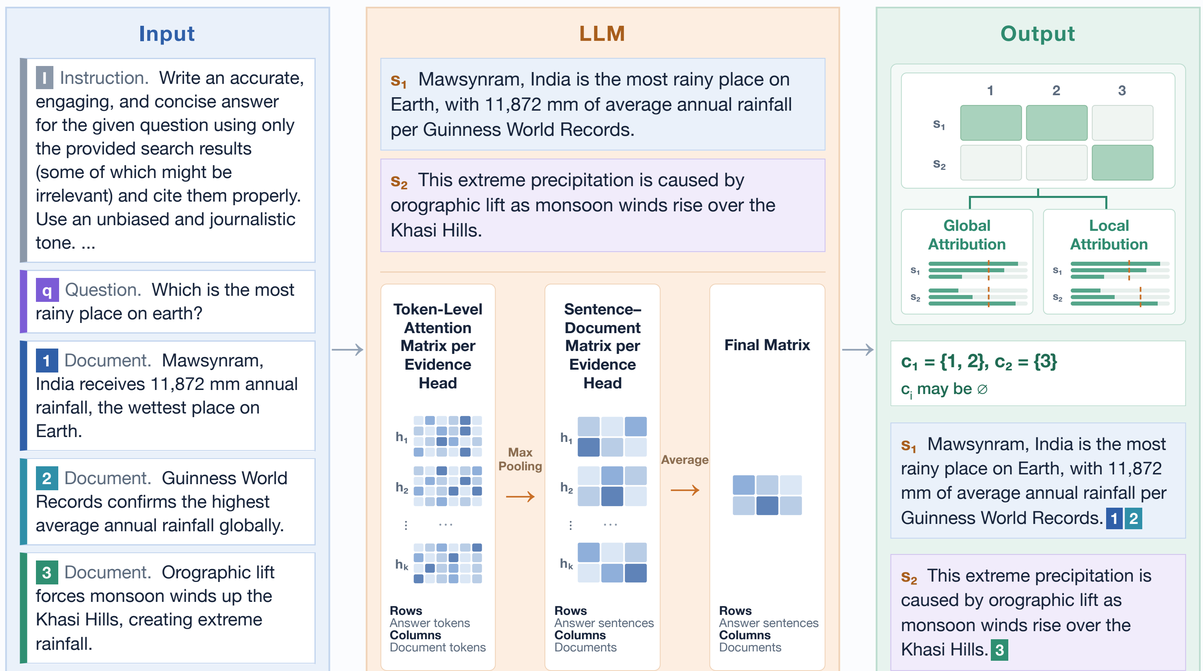
2 Evidence Head Selection

Q Do specialized attention heads responsible for evidence identification exist, and can they serve as a latent mechanism for answer attribution?



3 Attention-Based Attribution

Q Can attribution be derived directly from the internal mechanisms of LLMs, rather than being imposed externally?



4 Main Results

Q Does reading attention improve attribution over prompting and instruction tuning?

Methods	ASQA	ELI5	2WikiMQA	NQ	PopQA	TriviaQA	Overall
Llama-3.1-8B-Instruct							
Vanilla	67.49	49.10	31.53	61.31	64.68	53.05	54.41
Summary	65.21	45.32	30.46	57.72	60.27	52.78	51.85
Snippet	66.26	48.29	29.98	58.91	61.34	49.60	52.32
MIRAGE	66.22	47.41	30.30	61.69	64.86	53.63	53.91
C^2-Cite	67.82	49.19	32.25	61.62	65.09	53.94	54.87
Ours (global)	67.73	**51.41**	**34.75**	66.88	67.05	56.19	57.24
Ours (local)	**71.04**	51.11	32.39	**69.18**	**67.97**	**58.39**	**58.24**
Mistral-7B-Instruct-v0.3							
Vanilla	23.27	40.54	14.17	39.01	37.76	33.11	31.38
Summary	20.32	39.79	14.69	39.57	32.92	33.38	30.03
Snippet	28.18	36.97	18.91	36.53	38.95	35.83	32.45
MIRAGE	40.69	44.45	21.06	54.45	48.92	46.33	42.57
C^2-Cite	46.82	45.32	22.87	58.87	49.03	46.02	44.72
Ours (global)	64.32	**52.60**	**25.48**	65.70	60.05	52.44	53.34
Ours (local)	**65.51**	51.69	24.78	**66.68**	**60.58**	**53.65**	**53.72**
Qwen-2.5-7B-Instruct							
Vanilla	59.90	40.09	27.12	58.79	54.66	48.74	48.11
Summary	60.07	44.30	25.45	57.13	52.78	44.61	47.28
Snippet	61.22	43.73	26.26	57.05	56.54	46.48	48.44
MIRAGE	60.66	40.88	27.11	59.51	55.40	49.09	48.66
C^2-Cite	60.56	41.06	27.82	58.95	54.96	49.40	48.68
Ours (global)	64.22	44.90	**28.42**	65.82	61.12	53.80	52.95
Ours (local)	**68.46**	**46.69**	27.31	**68.40**	**65.02**	**55.14**	**55.05**

Local attribution records the best Overall F1 for all three model families, exceeding the strongest baseline by 3.37, 9.00, and 6.37 points.

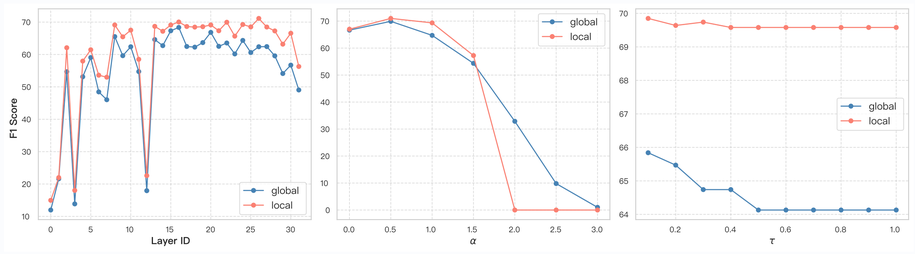
5 Ablation Study

Q Does the method rely on citation markers generated by the model?

Condition	ASQA	ELI5	2WikiMQA	NQ	PopQA	TriviaQA	Overall
Global							
Original	64.22	44.90	28.42	65.82	61.12	53.80	52.95
Citation markers removed	66.01	46.95	30.16	67.02	62.74	54.37	54.44
No citation prompt	61.17	40.65	27.40	59.91	57.24	50.34	49.35
Local							
Original	68.46	46.69	27.31	68.40	65.02	55.14	55.05
Citation markers removed	69.70	47.54	28.72	68.50	65.93	55.66	55.89
No citation prompt	64.04	39.47	25.72	62.23	59.81	51.89	50.41

Removing generated citation markers slightly improves Overall F1, while removing the citation prompt lowers scores but leaves a usable attribution signal.

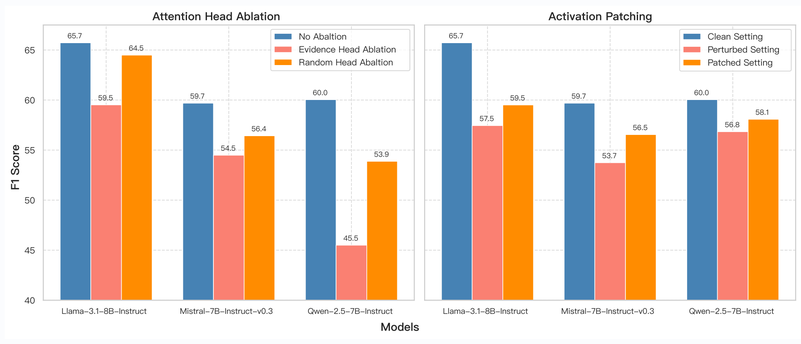
Q How does attribution change across layers, thresholds, and pooling choices?



Layer scores peak around layers 15 to 25, thresholds remain stable for alpha from zero to one, and attribution declines as pooling includes more tokens.

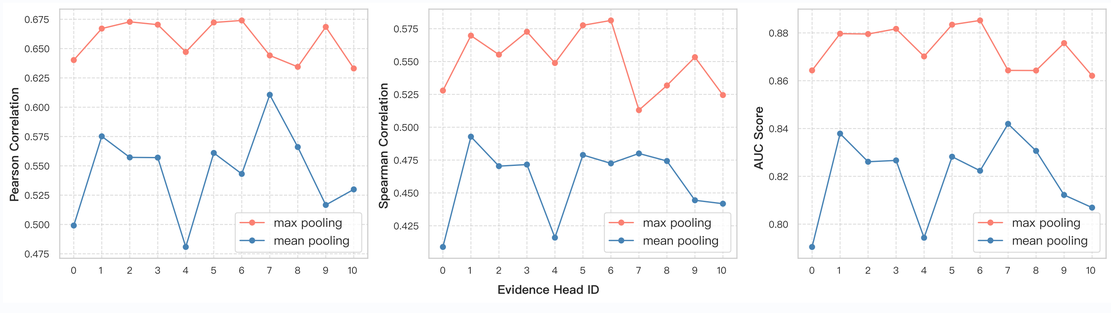
6 Causal Role and Semantic Alignment

Q How do Evidence Heads affect attribution under causal interventions?



Ablating Evidence Heads lowers attribution F1 more than ablating the same number of random heads in all three models, while activation patching partly restores performance.

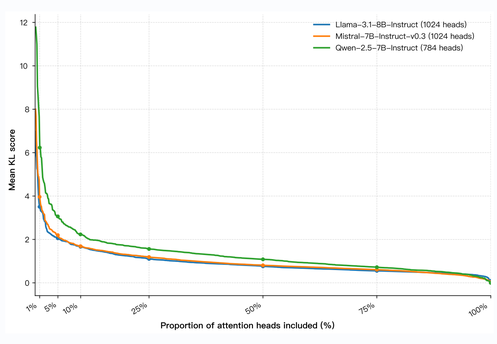
Q How closely do their attention patterns align with semantic support?



Max pooled attention aligns with NLI entailment better than mean pooling on Pearson, Spearman, and AUC, with lower variation across Evidence Heads.

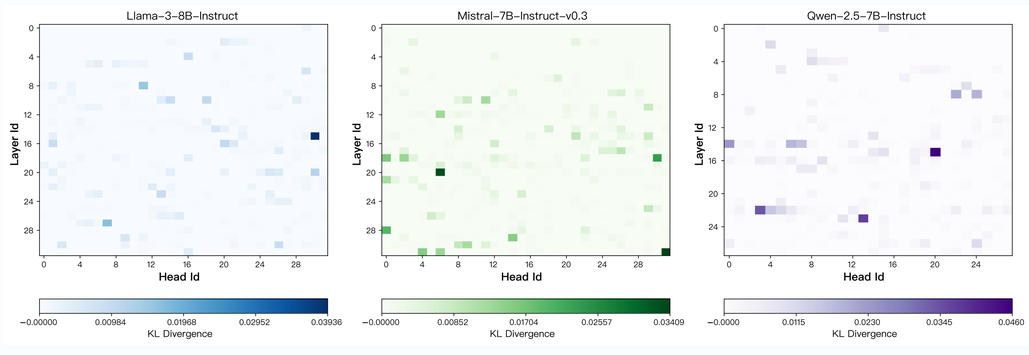
7 Evidence Head Count and Location

Q How is evidence sensitivity distributed across attention heads?



Mean KL scores fall sharply among leading heads before forming a long tail of low sensitivity heads, supporting sparse selection of the top 1 percent.

Q Where are Evidence Heads located inside each model?



Selected Evidence Heads occupy different positions in each model and cluster in intermediate layers, consistent with the layer attribution results.

Identifying and validating evidence sensitive attention heads in large language models and using them for training free sentence level answer attribution.