Attr-Kit: Efficient Toolkit for No-Decode Source Attribution

Sai Sundaresan*, Archit Gupta* , Debabrata Mahapatra
Adobe Research, India

In-context Attribution

Attributing agent outputs to the context for post-hoc verification

Auto-regressive (AR) citation generation:
*Additional decoding → **latency + cost***

Attribution is already encoded in model internals
Recover source influence directly from activations

Value Flow: Beyond where the model attends
Measure how source information contributes to the output direction

Attr-Kit: Efficient, no-decode attribution
Recover source attribution from early layers
Training-free • 0 decode tokens • Early exit

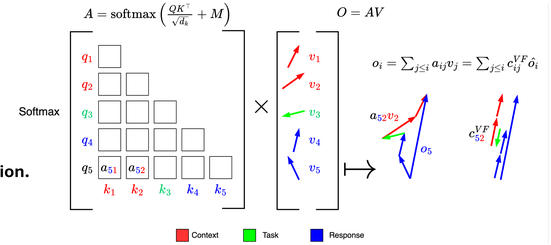
Where should I go to attend COLM?

Context:
[S1] COLM brings together researchers working on language modeling.
[S2] Previous editions of COLM were held in Philadelphia and Montreal.
[S3] COLM 2026 will be held in San Francisco, California.
[S4] The conference program includes invited talks, posters, and workshops.

This year COLM is in **San Francisco.[S3]**

Value Flow (VF): From Attention to Source Contribution

From attention to contribution:
VF combines attention with the geometric alignment between each source value and the attention-output direction, capturing **whether the attended information actually contributes to the target representation.**



VF vs. Other Attribution Signals

Token-Level Attribution Signals

Intuition	Signal	c_{ij}
Raw attention affinity	QK	$\frac{q_i^\top k_j}{\sqrt{d}}$
Attention weight	Attn	a_{ij}
Weighted Value magnitude	MV	$a_{ij}\|v_j\|$
Relative Value magnitude	MVN	$\frac{a_{ij}\|v_j\|}{\|o_i\|}$
Value contribution along output	VF	$a_{ij}(\hat{o}_i^\top v_j)$
Relative Value contribution	VFN	$\frac{a_{ij}(o_i^\top v_j)}{\|o_i\|^2}$

Pooling Strategies

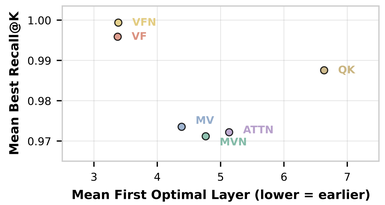
Purpose	Pool	C_{gs}
Preserve total signal	SS	$\sum_{i\in g}\sum_{j\in s} c_{ij}$
Normalize for source length	SM	$\sum_{i\in g}\frac{1}{\|s\|}\sum_{j\in s} c_{ij}$
Normalize for target length	MS	$\frac{1}{\|g\|}\sum_{i\in g}\sum_{j\in s} c_{ij}$
Normalize for both lengths	MM	$\frac{1}{\|g\|\|s\|}\sum_{i\in g}\sum_{j\in s} c_{ij}$
Capture strongest token pair	Max	$\max_{\substack{i\in g\\ j\in s}} c_{ij}$

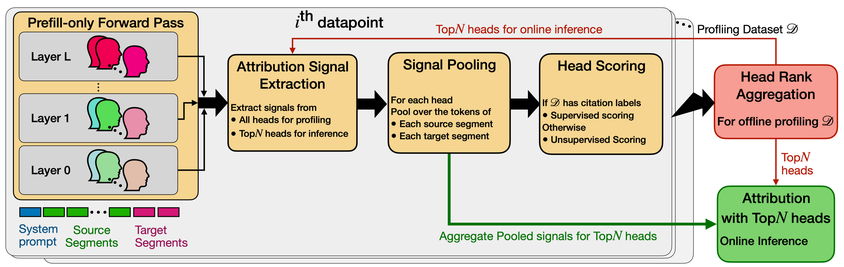
Token-Level Attribution → Segment-Level Attribution
Pooling aggregates token-pair signals over the tokens of each source and target segment.

How Early Do Attribution Signals Emerge? Analysis with Oracle:

For each example and signal, the oracle selects the best subset of attention heads for recovering the gold supporting sources.

VF & **VFN** reach near-perfect attribution in the earliest layers.





Offline Head Profiling With or Without Oracle Labels:

The oracle (gold citations) are not available at inference. Therefore, Attr-Kit does an offline head selection using a profiling dataset to select a fixed Top-N set to be used for online attribution.

Two Types of Profiling by Attr-kit depending on Oracle Label Availability

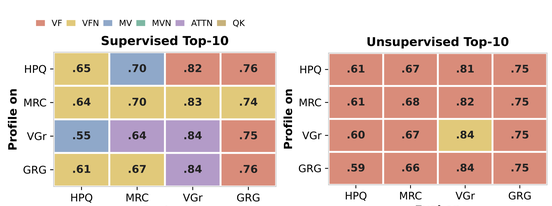
Head Profiling	Labels	Scoring Method	Definition
Supervised	Gold labels	Recall@K	$\text{score}_h^{(i)} = \frac{1}{G}\sum_g \text{Recall@K}\left(\text{signal}_h^{\text{seg},i}(g,\cdot), \text{gold}_g\right)$
Unsupervised	No labels	Signal Aggregation	$\text{score}_h^{(i)} = \sum_g \sum_s \text{signal}_h^{\text{seg},i}(g,s)$

Does VF Generalize?

Profile on one dataset → Evaluate on another

VF dominates 15 / 16 unsupervised transfers

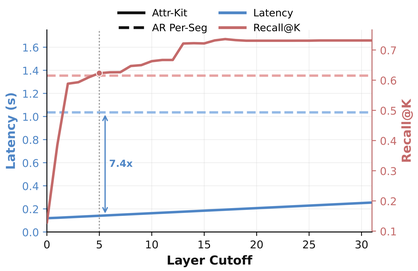
Datasets:
HPQ — HotpotQA
MRC — MultiRC
VGr — VeriGran · Contextual QA
GRG — GaRAGe · RAG grounding



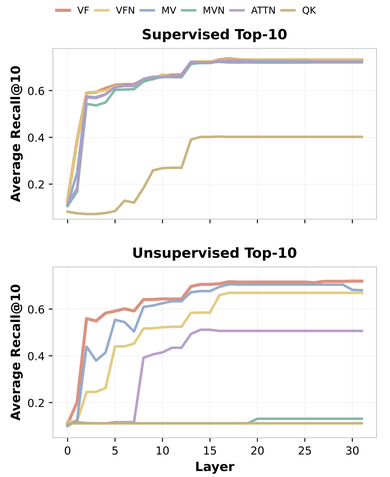
How Early Can We Attribute?

Stop at each layer → Measure attribution Recall@K* (number of gold docs)

VF reaches high recall earliest
and remains on the strongest accuracy–depth frontier.



7.4× Faster at Layer 5
*while exceeding same-model AR Recall@K**



Attr-kit vs others

Can No-Decode Attribution Beat AR Citation?
→ Yes — +11–25 pp across open-weight models

How Does Attr-Kit Compare? → Beats same-model AR, AT2[1], and ContextCite[2]

89% of frontier AR performance with an 8B model
*0.733 vs. 0.824 average Recall@K**.

Model	Method	HPQ	MRC	VGr	GRG	Avg
Qwen3-8B	Attr-Kit (VF)	0.625	**0.699**	**0.838**	**0.750**	**0.728**
	AR Per-Seg	**0.651**	**0.701**	0.628	0.488	0.617
	ContextCite	0.404	0.310	0.777	0.702	0.548
Llama-3.1-8B	Attr-Kit (VF)	**0.648**	**0.699**	**0.848**	**0.766**	**0.740**
	AR Per-Seg	0.637	**0.700**	0.631	0.422	0.597
	AT2	0.509	0.561	0.841	0.747	0.665
	ContextCite	0.534	0.355	0.770	0.707	0.591
Retrieval baselines						
	BM25	0.598	0.510	0.810	0.657	0.644
	Qwen3-Embedding-8B	0.672	0.483	0.690	0.433	0.570
Frontier non-AR references						
	OpenAI Emb.-3-Large	0.634	0.537	0.723	0.529	0.606
	Claude Opus 4.6 Citations	0.665	0.498	0.844	0.621	0.657
Frontier autoregressive						
	GPT-5.4 AR Per-Seg	**0.879**	**0.787**	0.864	**0.766**	**0.824**
	Claude Opus 4.6 AR Per-Seg	**0.878**	**0.785**	**0.871**	**0.763**	**0.824**

[1] AT2: B. Cohen-Wang, Y.-S. Chuang, and A. Madry. *Learning to Attribute with Attention.* arXiv:2504.13752, 2025
[2] ContextCite: Cohen-Wang et al. *ContextCite: Attributing Model Generation to Context.* NeurIPS, 2024.