When Graph Tokens Sink: A Mechanistic Analysis of Graph Language Models

Ding Zhang[1], Runtao Zhou[1], Wenqing Zheng[2], Rizal Fathony[2], Bayan Bruss[2], Chirag Agarwal[1]

School of Data Science, University of Virginia[1]
Capital One[2]

1 Research Importance & Motivation

- Graph Language Models (GLMs) flatten graph topology and node features into graph tokens so an LLM can read graphs alongside text instructions.
- Transformers are highly prone to structural pathologies when processing sequential data.
- Transformers develop attention sinks and massive activations: a few tokens become internally "loud" for architectural reasons, not because they carry meaning.
- **Central question: when a graph token becomes 'loud' within the LLM, is the model actually using it for topological reasoning, or is it just an architectural artifact?**

2 Setup: Detecting Graph Sink Tokens

- **Definition.** A graph token j is a graph sink token at layer l if the RMS-normalized magnitude of its hidden state on a small set of sink dimensions exceeds a threshold τ, namely:

$$\widehat{\mathcal{I}}_g^l = \{ j \in \mathcal{I}_g \mid \phi(\mathbf{x}_j^{l-1}) \geq \tau \},$$

where, $\phi(\mathbf{x}_j^{l-1}) = \max_{\bar{d} \in \mathcal{D}_{\text{sink}}} \left| \text{RMSNorm}(\mathbf{x}_j^{l-1})_{\bar{d}} \right|.$

- **LLaGA** — node-aligned graph tokens. Center node at index 0, k-hop neighbors in a fixed tree template, [PAD] fills unused slots (K = 111 / 222).
- **TEA-GLM** — a GNN encoder compresses the neighborhood into K = 5 learned graph tokens through a linear projector.
- **Benchmarks:** Cora, Arxiv, PubMed — node classification and link prediction tasks, with LLaMA-family (Vicuna-7B) backbones.

3 Finding 1: Sink Tokens Are Activation Outliers

- Across datasets, tasks, and both architectures, a few hidden dimensions carry huge spikes. Dimension 2533 is a known LLaMA sink dimension, but **dimension 1512 appears only after graph information is mapped into the token space.**

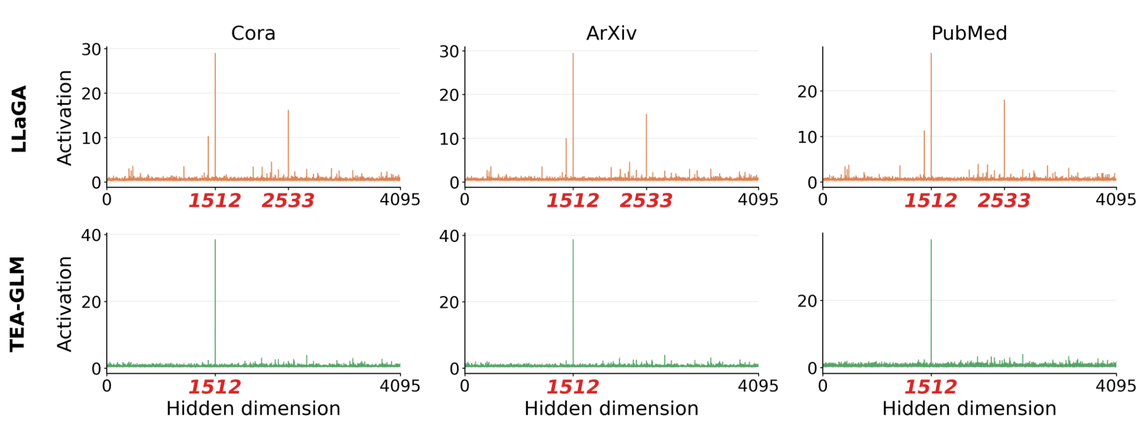

Figure 1. Average activation magnitude of detected graph sink tokens for node classification task. A small set of hidden dimensions (1512, 2533) forms large spikes while all others stay flat across both architectures.

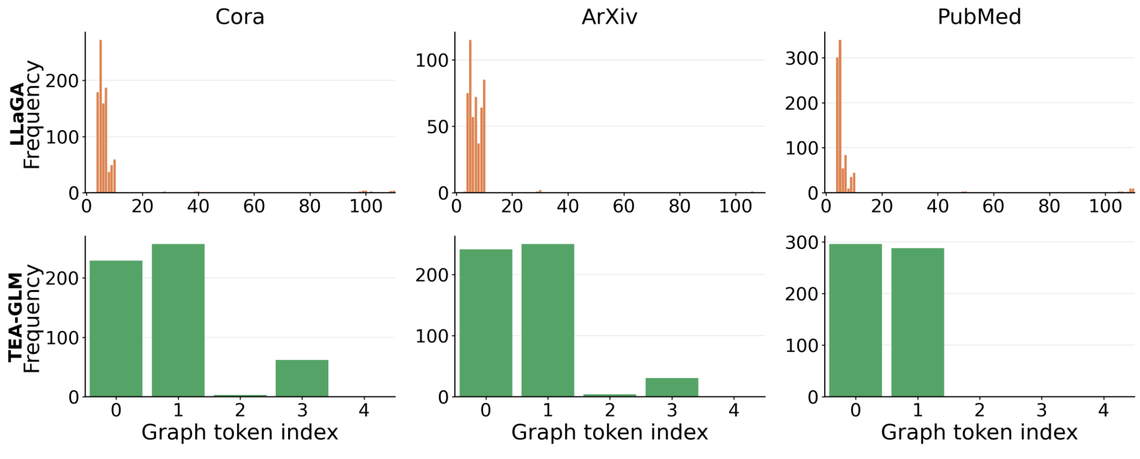

Figure 2. Positions of detected sink tokens for both architectures for node classification task. Sinks concentrate at early graph-token indices. In LLaGA, the top-2 sinks are always [PAD] tokens.

4 Finding 2: Loud ≠ Attended

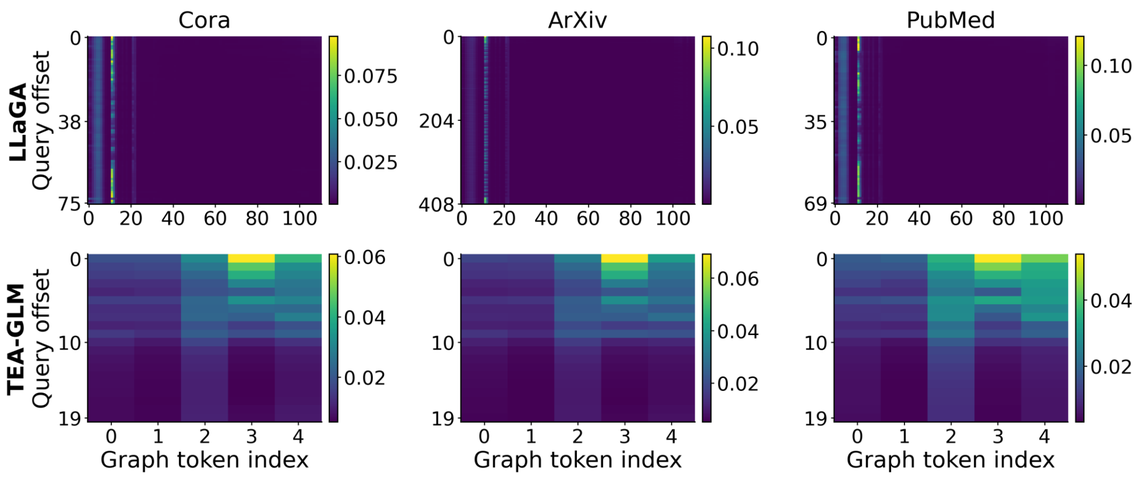

Figure 3. Query-to-graph attention maps for the node classification task across three benchmark datasets and on two GLM architectures, averaged over heads and samples.

TEA-GLM sinks sit at positions 0–1, yet query attention is often strongest on positions 2–4 — across query offsets and across layers. LLaGA shows stable attention bands, but they are not exclusive to sink tokens, and its top sinks are [PAD]. **Activation saliency does not translate into attention dominance**, so graph sink tokens are not simply graph-domain versions of classical attention sinks.

5 Finding 3: Sinks Are Not the Information Carriers

- Setup: we prune, swap, and reposition graph sink tokens and measure the damage. Removing the top-2 sink tokens barely moves accuracy; **removing two random non-sink tokens often hurts more** (red cells below).

Model	Intervention	Arxiv	Cora	PubMed
LLaGA	Baseline	77.00	88.40	94.60
	Top-2 Sink	77.00	88.00	95.00
	Non-sink	**74.36 ± 1.44**	**80.48 ± 2.47**	**89.16 ± 4.58**
	Swap	76.64 ± 0.27	87.74 ± 0.39	94.12 ± 0.27
TEA-GLM	Baseline	56.67	13.33	83.67
	Top-2 Sink	56.67	13.00	83.67
	Non-sink	**44.40 ± 1.13**	**12.80 ± 1.72**	83.40 ± 1.09
	Swap	56.40 ± 0.43	12.87 ± 0.18	83.27 ± 0.43

Table 1. Node classification accuracy under graph-token interventions. Pruning the two most salient sink tokens changes almost nothing; pruning random non-sink tokens (15 seeds) drops accuracy more.

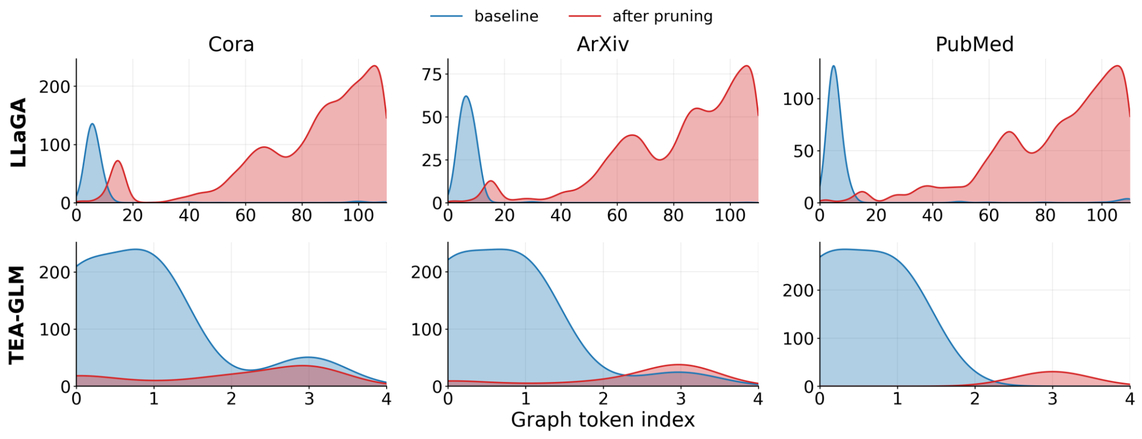

Figure 4. Sink positions before (blue) and after (red) pruning all sinks. LLaGA re-creates sinks spread across many positions; TEA-GLM activations drop and sinks mostly disappear.

- Across pruning, swapping, repositioning, and post-pruning sink detection experiments, graph sink tokens do not behave as the main semantic or structural carriers in graph language models.
- After pruning, LLaGA redistributes sink behavior across many graph-token positions, whereas TEA-GLM largely loses sink-qualified activations. These results connect back to our central observation: activation-level saliency does not imply graph-information utility.

6 Finding 4: Sink States Carry Weak Semantics

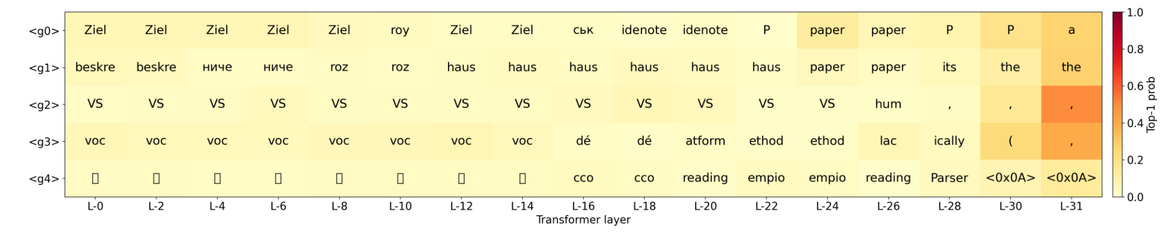

Figure 5. Logit lens on TEA-GLM graph tokens on Pubmed node classification task. Each cell is the most frequent top-1 decoded word per position and layer; color is its average probability.

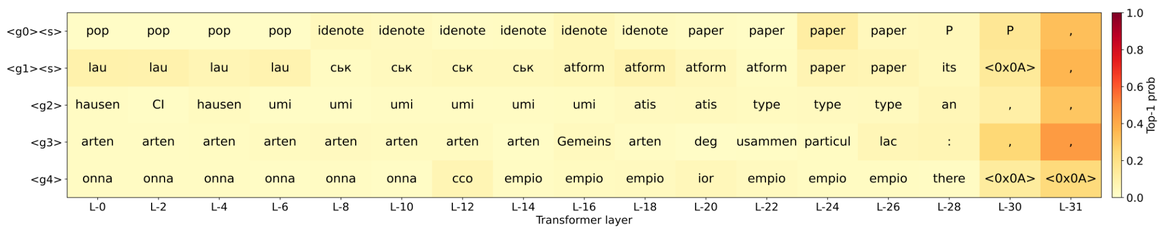

Figure 6. Logit lens on TEA-GLM graph tokens on Arxiv node classification task. Each cell is the most frequent top-1 decoded word per position and layer; color is its average probability.

- Decoding graph-token hidden states through the LM head yields fragmented subwords and generic terms with low probability. From layer ~20, sink positions g0 and g1 repeatedly decode to **"paper"** on both datasets: a broad domain prior from pretraining, not class labels or topology.
- Thus, logit lens analysis shows that graph sink tokens mainly expose generic domain-level terms rather than task-specific or topology-aware graph information.

Conclusions & Discussions

- **Activation-level saliency of graph tokens is decoupled from graph-semantic utility.**
- Mapping graph structure into the LLM token space does not, by itself, produce a usable topology-aware internal representation.
- Diagnosing GLMs by internal magnitude alone can be misleading; interventions and decoding checks are needed.
- Limitations and Future Work:
  - Other GLM architectures may exhibit different graph sink token behaviours.
  - Future work should include: better graph-token construction, placement, and graph-token alignment, to better preserve topology-aware information inside LLMs.