How Graphs Ground Large Language Models – Counterfactuals for Subgraph Verbalizations

Sara Buchmann, Emanuel Slany, Stephan Scheele

Research Project Data Tales, funded by the Bavarian StMWi (DIK0660)

Data Tales

GraphRAG improves **factual accuracy** by grounding LLMs in **knowledge graphs**.

*But which nodes are actually **relevant**?* Individual node contributions to the output remain **unclear**.

We aim to quantify how **individual nodes influence** LLM outputs and provide **explanations** of their effects.

Input: query q + subgraph G with nodes V, retrieved via GraphRAG

Performing **node occlusion** by removing individual nodes $n \in V \rightarrow \alpha(G, \{n\})$

LLM generates original/perturbed **answer**:

$$a = \phi(G, q), \quad a_n = \phi(\alpha(G, \{n\}), q)$$

Outputs are mapped to **embeddings** $\psi(a), \psi(a_n)$

Overview of the Method

Measuring the resulting **semantic change** Δ using a **dissimilarity** function d

$$\Delta(\{n\}; G, q) := d(\psi(a), \psi(a_n))$$

Node Relevance Distribution for (G, q) with

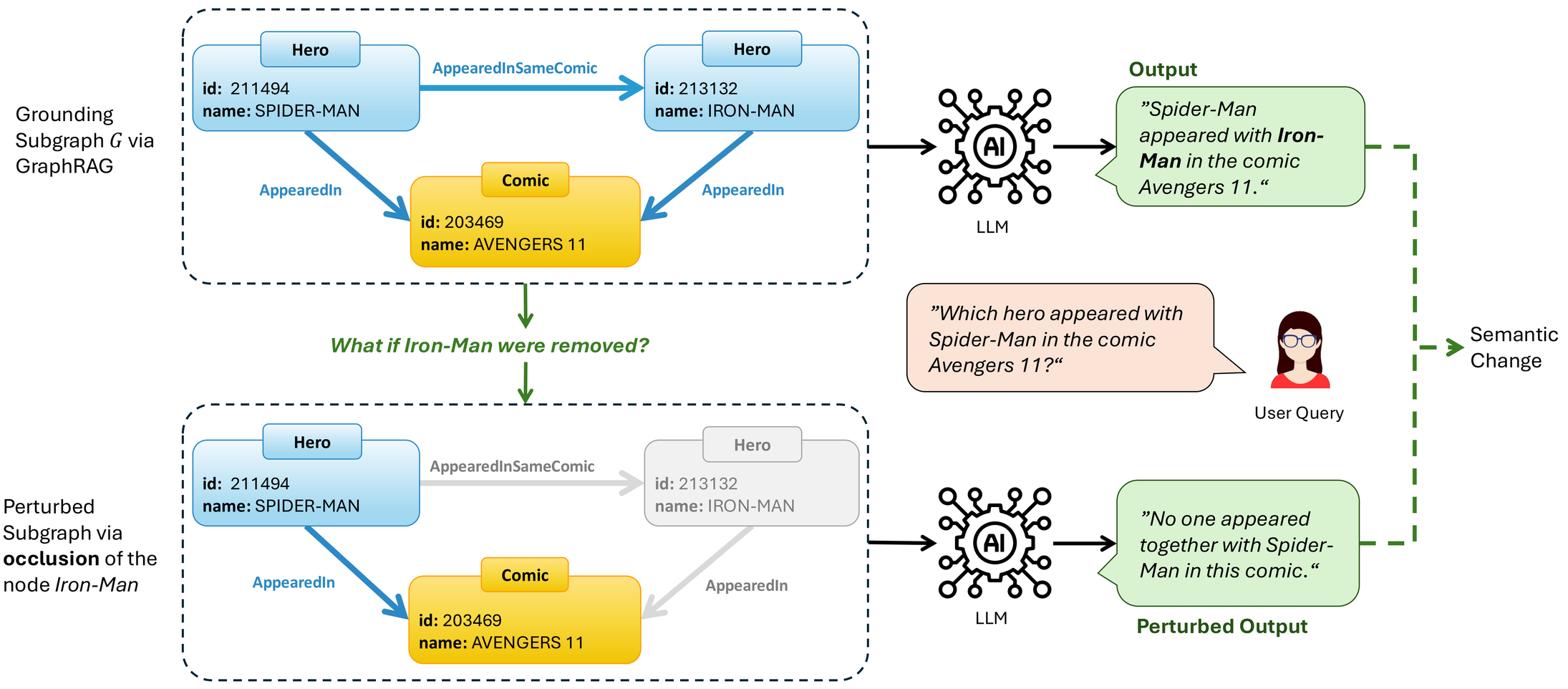
$$\delta(n; G, q) := \Delta(\{n\}; G, q) \quad \forall\, n \in V,$$

$$\pi(n; G, q) := \begin{cases} \frac{\delta(n;G,q)}{\sum_{m \in V} \delta(m;G,,q)}, & if \ \sum_{m \in V} \delta(m; G, q) > 0 \\ \frac{1}{|V|}, & otherwise. \end{cases}$$

Minimal counterfactual explanations are identified using **greedy** or exhaustive **search**

Counterfactual Explanation of (G, q) for a dissimilarity threshold $\tau > 0$, is

$$N^* \in \arg\min_{N \subseteq V} |N| \text{ subject to } \Delta(N; G, q) \geq \tau.$$



Evaluation and Results

RQ1: ***How stable are node relevance values under different LLM runs?***

RS: stability of node relevance value rankings across LLM runs
RS_1: stability of the top-ranked nodes
W: Kendall's concordance

- Top-ranked nodes are highly stable
- Lower-ranked nodes show lower stability and increasing variability
- Agreement of node rankings within a graph decreases with graph size

RQ2: ***Are node relevance values approximately additive?***

MAE: deviation between observed $\Delta(N; G', q)$ and predicted $\sum_{n \in N} \delta(n; G', q), \ N \subseteq V'$ semantic change

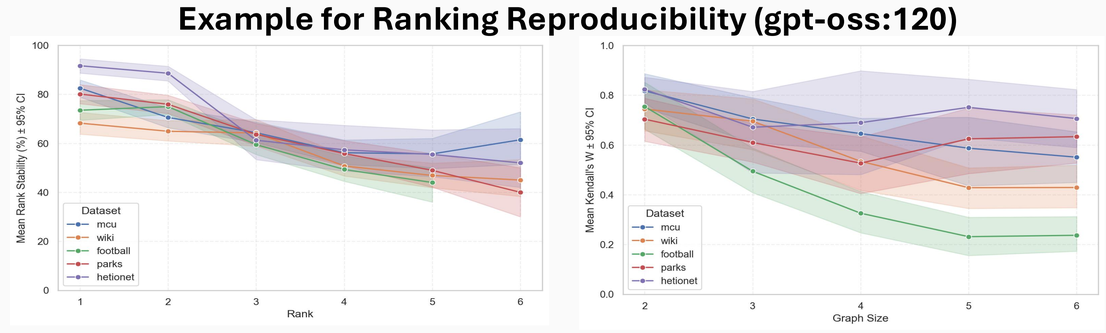
- Additivity holds approximately across datasets and models
- Only few cases showing interaction effects between nodes

RQ3: ***Can minimal node subsets be identified that determine the output?***

CR_1: how often greedy finds a minimal solution
SR: fraction of single-node explanations

- Greedy search is near-optimal in most cases
- Often a single node is sufficient to determine the output semantics

Example for Ranking Reproducibility (gpt-oss:120)



Dataset	Model	$\overline{RS}$	$\overline{RS_1}$	$\overline{W}$	$\overline{MAE}$	CR_1	SR
football	gemma4:e4b	62.75 ± 21.02	67.77 ± 23.29	0.43 ± 0.30	0.22 ± 0.13	**0.99**	**0.96**
	gpt-oss:120b	68.97 ± 21.84	73.49 ± 23.54	0.43 ± 0.28	0.12 ± 0.11	0.77	0.86
	llama3:70b	**80.40 ± 19.37**	**83.68 ± 19.99**	0.36 ± 0.29	0.12 ± 0.12	0.74	0.70
	qwen2.5:72b	79.58 ± 19.57	83.07 ± 19.21	**0.45 ± 0.35**	**0.06 ± 0.05**	0.51	0.80
hetionet	gemma4:e4b	66.34 ± 16.84	66.40 ± 17.66	0.26 ± 0.32	0.30 ± 0.19	**1.0**	**1.0**
	gpt-oss:120b	83.31 ± 21.07	91.65 ± 15.28	0.79 ± 0.23	0.18 ± 0.10	0.97	**1.0**
	llama3:70b	**89.30 ± 15.32**	**94.90 ± 10.22**	**0.81 ± 0.19**	0.15 ± 0.08	0.90	0.99
	qwen2.5:72b	77.20 ± 18.64	75.78 ± 17.13	0.26 ± 0.37	**0.14 ± 0.06**	0.90	0.97
mcu	gemma4:e4b	63.52 ± 22.30	72.05 ± 23.25	0.53 ± 0.33	0.19 ± 0.14	**0.95**	**0.97**
	gpt-oss:120b	69.14 ± 21.92	82.50 ± 19.35	0.67 ± 0.23	0.10 ± 0.07	0.93	0.88
	llama3:70b	73.92 ± 21.58	**83.51 ± 20.17**	**0.70 ± 0.24**	0.16 ± 0.13	0.82	0.84
	qwen2.5:72b	**77.26 ± 23.43**	82.74 ± 22.21	0.66 ± 0.36	**0.08 ± 0.05**	0.84	0.85
parks	gemma4:e4b	65.89 ± 19.65	70.00 ± 20.04	0.46 ± 0.32	0.27 ± 0.18	**1.0**	**0.94**
	gpt-oss:120b	70.80 ± 21.31	80.09 ± 20.07	**0.63 ± 0.24**	0.15 ± 0.13	0.96	0.93
	llama3:70b	79.51 ± 19.59	**89.62 ± 15.06**	0.61 ± 0.27	0.10 ± 0.08	0.75	0.85
	qwen2.5:72b	**82.55 ± 19.39**	84.04 ± 17.82	0.49 ± 0.41	**0.08 ± 0.05**	0.73	0.86
wiki	gemma4:e4b	66.21 ± 21.80	72.73 ± 21.25	0.57 ± 0.33	0.17 ± 0.11	**0.97**	**0.99**
	gpt-oss:120b	61.98 ± 22.97	68.22 ± 24.41	0.58 ± 0.26	0.15 ± 0.09	0.94	**0.99**
	llama3:70b	72.40 ± 19.96	80.36 ± 20.00	0.64 ± 0.20	0.13 ± 0.13	0.90	0.89
	qwen2.5:72b	**79.11 ± 22.25**	**84.68 ± 20.34**	**0.70 ± 0.36**	**0.10 ± 0.06**	0.87	0.86

Model comparison across datasets

$\overline{RS}$: mean rank stability (all nodes);

$\overline{RS_1}$: mean rank stability for first-ranked nodes;

$\overline{W}$: mean Kendall's concordance;

$\overline{MAE}$: mean absolute error of relevance scores;

CR_1: minimality rate;

SR: sufficiency rate.

Key Findings

❖ Important nodes are consistently identified across LLM runs
→ node rankings are reproducible, especially for top-ranked nodes

❖ Semantic influence is highly concentrated on few nodes
→ a small subset of nodes accounts for most output changes

❖ Minimal counterfactual explanations are frequently achieved by a single node → greedy search finds near-optimal solutions, often of size one

Limitations and Future Work

– Explanations rely on node interventions limited to occlusion
→ *explore richer interventions (replacement, perturbation)*

– A fixed threshold τ may not reflect task-specific sensitivity to semantic change
→ *adapt thresholds based on task and data*

– Evaluation is based on embedding similarity, not human judgment
→ *validate explanation quality via user study*