SOVEREIGN: What is the relationship between model size (e.g., 7B vs 70B parameters) and the transferability of token-leve

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Graph Neural Networks (GNNs), specifically designed to process the graph data, have achieved remarkable success in various applications. Link stealing attacks on graph data pose a significant privacy threat, as attackers aim to extract sensitive relationships between nodes (entities), potentially leading to academic misconduct, fraudulent transactions, or other malicious activities. Previous studies have primarily focused on single datasets and did not explore cross-dataset attacks, let alone attacks that leverage the combined knowledge of multiple attackers. However, we find that an attacker

1 Introduction

Analysis of: Large Language Models Merging for Enhancing the Link Stealing Attack on Graph Neural Networks. Research goal: What is the relationship between model size (e.g., 7B vs 70B parameters) and the transferability of token-level adversarial attacks across SOTA legal reasoning models on the Abductive Logic Reasoning (ALR) benchmark, evaluated via accuracy degradation?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

11 papers retrieved. 9 claims extracted, 0 verified. Tribunal: 2.3/10 → REJECT (revision_round=0). Policy: ESCALATE_TO_OWNER.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
Large language models require substantial computational resources, making them challenging to manage in practical applic	×	0.07
In link stealing attacks, attackers infer the existence of links between nodes by leveraging node knowledge and response	×	0.13
Model merging combines two or more trained models into a single model without requiring data sharing, thus safeguarding	×	0.08
The merged model from multiple attackers has stronger generalization capabilities compared to a single attacker model.	×	0.14
The link stealing attack process involves sending node features to the target model to obtain posterior probabilities.	×	0.11
The merged model achieves accuracy scores of 0.94, 0.91, and 0.83 on Dataset 1, Dataset 2, and Dataset 3 respectively.	×	0.05
The merged model achieves F1 scores of 0.90, 0.92, and 0.86 on Dataset 1, Dataset 2, and Dataset 3 respectively.	×	0.05
The proposed LLM-based method achieves a mean accuracy of 93.79±0.04 across datasets.	×	0.05
The proposed LLM-based method achieves a mean F1 score of 91.73±0.04 across datasets.	×	0.04

References

- http://arxiv.org/abs/2403.09832v1
- http://arxiv.org/abs/2304.06912v2
- http://arxiv.org/abs/2412.05830v1