Report: Enhancing Knowledge Graph Attacks via Multi-Agent Abstract Generation

In this project, we investigate the vulnerability of biomedical knowledge graphs (KGs) to adversarial textual inputs by leveraging large language models (LLMs). Our starting point is *Scorpius*, a KG poisoning framework that employs a single model (BioGPT) to generate malicious biomedical abstracts. While effective, this single-model setup limits the diversity and adaptability of the generated content. To address this limitation, we propose a multi-agent abstract generation strategy utilizing both Gemini and Mistral. By introducing variation in abstract styles and linguistic features, we aim to enhance the stealth and success rate of KG poisoning attacks.

We begin by selecting a small, representative subset of five drug–disease pairs from the GNBR dataset (Global Network of Biomedical Relationships), which is constructed from PubMed abstracts and organized into triples of the form (Drug, Relation, Disease). Using PubTator and GNBR's sentence-extraction tools, we prepared the relevant textual data for manipulation.

Our methodology consists of several stages. First, we prepare the selected pairs and construct prompts (Step 1). In Step 2, we employ Gemini and Mistral to generate fake but scientifically plausible abstracts based on these prompts. The generated abstracts are saved and prepared for injection. In Step 3, we execute the attack.py script to compute adversarial edges, identifying which triples, if introduced, would most significantly affect KG reasoning. Step 4 involves injecting the generated abstracts into the KG using edge_to_abstract.py, replacing the original content with poisoned abstracts. Finally, in Step 5, we retrain the DistMult model on the modified KG and evaluate changes in ranking performance for the targeted drug–disease relationships.

Our evaluation indicates that Gemini-generated abstracts are highly effective in poisoning the KG. The Mean Reciprocal Rank (MRR) increased from 0.006 to 0.076, and Hits@10 rose from 0.000 to 0.200, while the Mean Rank dropped significantly from 220.2 to 82.6. These results suggest that Gemini-generated texts successfully boosted the ranking of the targeted triples. By contrast, Mistral's outputs were less impactful: MRR increased only to 0.019, and Hits@10 showed no improvement. This disparity implies that abstract quality and the linguistic characteristics of different LLMs play a crucial role in poisoning effectiveness.

Looking forward, we plan to scale up from 5 to 100 drug–disease pairs to simulate a more realistic attack scenario and persistent threats by repeatedly injecting new poisoned abstracts and retraining the KG model over time. This project highlights a significant vulnerability: biomedical KGs often rely on textual inputs without sufficient adversarial detection mechanisms. As LLMs continue to advance, the development of robust defenses against subtle, text-based poisoning becomes increasingly urgent.