Multi LLM Attacks on Biomedical KG

Overview

The **Scorpius framework** demonstrates how fake biomedical abstracts generated by LLMs can be used to poison biomedical knowledge graphs (KGs), impacting real-world drug discovery tools. However, the original framework used only **GPT-3.5**, whose repetitive style and limited linguistic diversity made it vulnerable to AI based detection.

This project explores whether leveraging **multiple modern LLMs** (Claude 3 Opus and GPT-4o) can improve the stealth and effectiveness of such attacks, by generating more stylistically diverse fake abstracts for drug disease pairs.

Methods

The project uses the **GNBR dataset** (Global Network of Biomedical Relationships), which contains drug-disease gene relationships extracted from **PubMed abstracts**. The KG is constructed from this data and used to train link prediction models.

Target drug-disease pairs were selected for the attack. For each pair, multiple fake abstracts were generated using **Claude 3 Opus** and **GPT-4o**. These abstracts were formatted to be **Scorpius-compatible** and injected into the training set to test whether they could boost the ranking of false triples in the KG.

Compared to the baseline (GPT-3.5-only), this multi-LLM approach aimed to:

- Increase **linguistic variance** and reduce detectability.
- Improve **poisoning efficacy** by providing a more realistic and varied set of fake abstracts.

Results

The project partially achieved its goal:

- The abstracts generated by Claude and GPT-4o were noticeably **more diverse** in tone and phrasing.
- Integrating these abstracts into Scorpius was more complex than expected (With more time would eliminate this problem).
- Without robust scoring and evaluation, it was difficult to measure which LLM produced more effective poisonings. But this is a good direction to work towards.

Overall, this suggests that **expanding beyond a single and more newer LLM is a promising direction** for improving stealth in KG attacks, though better integration and evaluation tools are needed.

Future Work

With more time, I would:

- Get more familiar with the Scorpius framework.

- Improve the automation and integration of LLM generated abstracts.

- Develop better scoring mechanics to compare abstract quality and poisoning impact.

- Explore additional LLMs (e.g., DeepSeek, Gemini) for more diversity.

Reflection

This project highlights both the **power and risks of LLMs** in scientific domains. Biomedical KGs are highly vulnerable to poisoning, and the ease of generating realistic fake data with LLMs raises serious concerns about **AI safety, trust, and responsibility** in medical research.