h e g

Haute école de gestion
Genève

21st June 2023

MONITORING LARGE GENERATIVE LANGUAGE MODELS:

TRENDS AND IMPACT ON CYBERSECURITY AND ON COMPETITIVE INTELLIGENCE

Hes·so

HES-SO Team

h e g
Haute école de gestion
Genève

Ciarán Bryce
Professeur HES

Hélène Madinier
Professeure HES

Alexandros Kalousis
Professeur HES

Thomas Pasche
Assistant HES

Ilan Leroux
Assistant HES

Patrick Ruch
Professeur HES et Responsable de la recherche

Objectives

h e g
Haute école de gestion
Genève

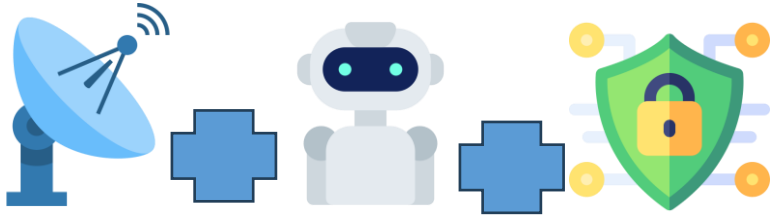
Monitoring trends and support the DDPS cyber defense **strategy**

- **Conducting a bibliometric** analysis of publications related to major **generative language models**
- **Tracking** the trends and cybersecurity implications of **large generative language models**
- **Tracking** the trends and competitive intelligence implications of **large generative language models**

Ambitions

h e g
Haute école de gestion
Genève

We intend to become an active center of Competence in



Competitive intelligence | **Artificial intelligence** | **Cybersecurity**

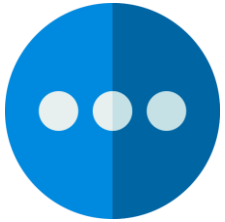
This means an active approach to technology watch

Workshops

Networking

Dedicated web page

Internal communication

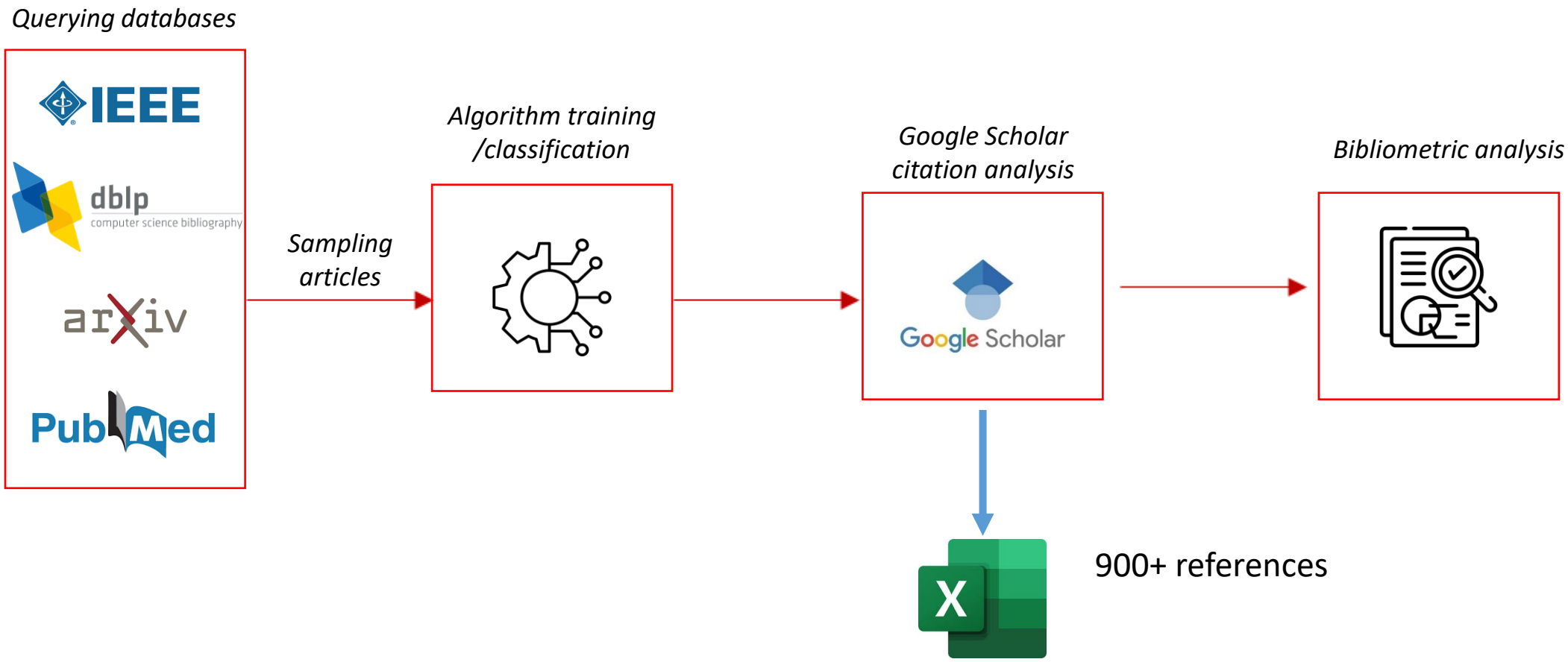


and more...

Methodology

Haute école de gestion
Genève

Data pipeline



Flowatcher in action

Haute école de gestion Genève

A software in constant evolution

Secure Code

- Repair cybersecurity bugs
  - Difficult without comprehensive sources
- Copilot uses Codex
  - Trained on public Github repos
- Code is not always robust to certain attacks
- Can introduce security bugs

Testing

- Red Teaming
- Adversarial Testing
  - Generation adversarial examples
- Mutation Testing
  - Make changes to input data to see effect
- Using regex to speed up testing (ReLM)
  - Succinct queries, scalable testing

Attacks

h e g
Haute école de gestion
Genève

- Direct and indirect prompt injection
  - Manipulating prompts to cause leakage, DoS or toxic content
  - Sydney chatbot replica created from hidden text (font size, color)

- Model poisoning - Inject malicious data into training data
- Adversarial examples E.g., fake news article
- Data poisoning (triggers) - Syntax-based and translation-based
- Backdoor attacks
  - Trigger for specific inputs
- Model inversion attacks
  - Reconstruct training data or input data from model's output
    - E.g., Samsung attack

Jailbreaking Strategies

- Attention shifting
  - Shift model's attention from Q&A to story telling
  - E.g., Text continuation

- Pretending
  - Pretend to have different goals to actual intentions (over 97% of jailbreak attacks)
  - E.g., Role-playing games

- Privilege escalation
  - Incite model to break imposed restrictions rather than by-passing them
  - E.g., sudo mode pattern

Shields

- Training on Dark Web corpus, e.g., DarkBERT
- Honeypots
- Privacy preserving prompt tuning
- Structured access schemes to control interaction
- AIGC detection - Watermarking, classifier-based, likelihood based
- Models
  - Generative Adversarial Network models for SPAM and anomaly detection
  - Generative Pre-trained Transformer models for NLP tasks, fake Cyberthreat intelligence
  - BERT-based models for SOC applications
- Threat modeling (STRIDE)
- SentinelOne and Microsoft Copilot
- ChatGPT for user authentication

Spears and Metrics

h e g
Haute école de gestion
Genève

- Phishing
  - Automating labor-intensive tasks
    - Finding targets, personalized messages (measured by coherency, grammar, ….)
    - *Replying to ransomware payment questions*

- Social engineering
  - Using ChatGPT adds to deliver malware (e.g., Redline Stealer malware)

- Malware
  - Zero-knowledge implementation of code for Top 10 Mitre Attacks
  - Polymorphic malware (e.g., Black Mamba keylogger)
  - From Stack Overflow questions to malicious libraries
  - Vulnerabilities in ChatGPT libraries

Societal Impact

h e g
Haute école de gestion
Genève

- Use of “regulatory sandbox” to facilitate experimentation in AI
  - G7 leaders calling for controls

- Impact of algorithms on poverty
  - World Bank-funded algorithm to determine which families should receive financial assistance
- Role of humans as data workers for AI
- Real risks of AI
  - Security, ethics and politics.
- Measurement and mitigation of bias and hostile AI

Competitive intelligence impact

h e g
Haute école de gestion
Genève

- LLM models : **virtual assistants** for CI pros
  - Especially helpful for the **analysis phases** of the competitive intelligence (summarizing, synthetizing, report generation)

- Multiplicity of **new tools** allowing to search for web information (webchatgpt extension, You, perplexity.ai, Bing chat...)

- New strategy for BING

  => many search engine who used the free BING API cannot use it anymore (Duckduckgo, Ecosia, Qwant, Brave Search etc...)

  => **Search engine rarefaction**

Thank you for your attention

h e g
Haute école de gestion
Genève

Your questions are welcome!