14. Emerging Topics - Adversarial Retrieval

Peter Carragher

Carnegie Mellon University

Adversarial Adaptation of Misinformation Sites

YourNewsWire.com

News. Truth. Unfiltered.

HOME | NEWS | HEALTH | SCI/ENVIRONMENT | TECHNOLOGY | ENTERTAINMENT

CONTACT US | TERMS OF USE | PRIVACY | ADVERTISE

HEADLINES > [June 1, 2018] FBI: 'Indisputable Evidence' That Obama Paid MI6 To Fake Trump Dossier ▸ NEWS

NEWS PUNCH

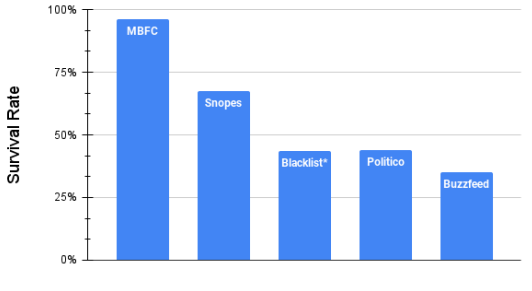
WHERE MAINSTREAM FEARS TO TREAD

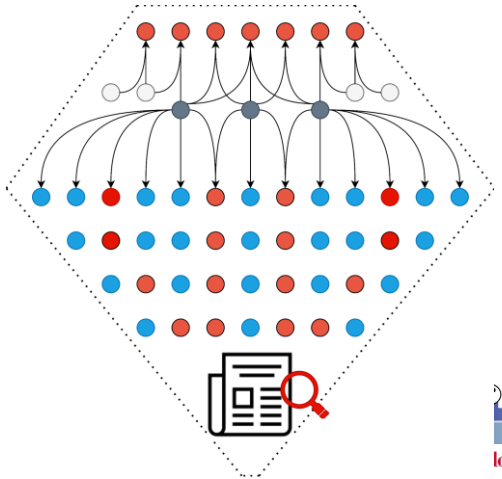
```
Loading...

    http://yournewswire.com/ |
    20:08:59 February 20, 2019

    Got an HTTP 301 response at crawl time

Redirecting to...

    https://newspunch.com/
```

HOME | NEWS | HEALTH | SCI/ENVIRONMENT | TECHNOLOGY | ENTERTAINMENT

CONTACT US | TERMS OF USE | PRIVACY | ADVERTISE

HEADLINES > [February 1, 2019] Jury Awards Sen. Rand Paul $580,000 to Be Paid by Antifa Thug Who Assaulted Him



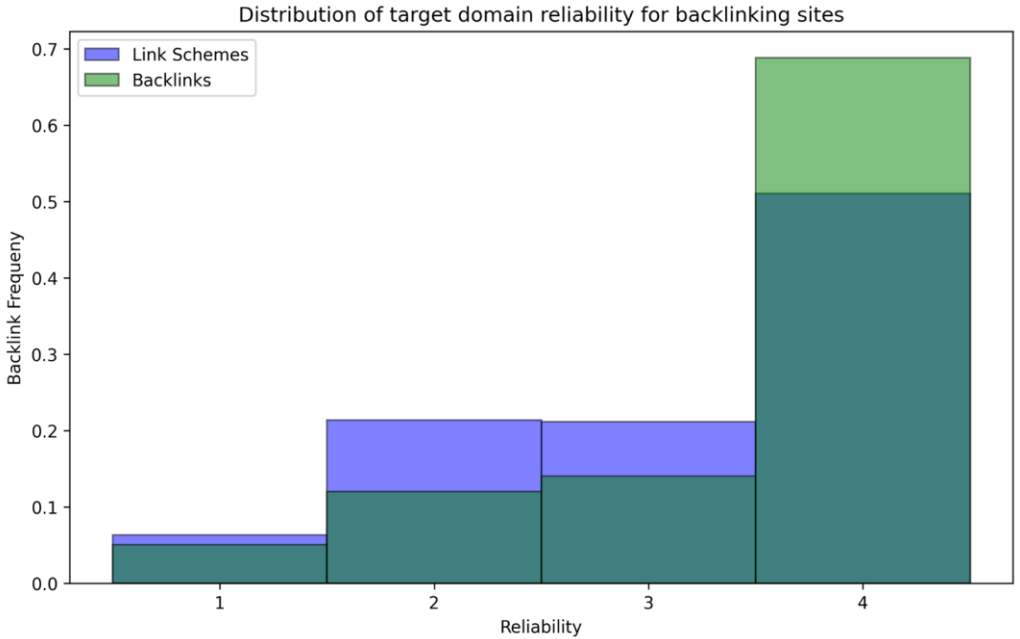


Peter Carragher, Evan M. Williams, Kathleen M. Carley. 2024.
"Detection and Discovery of Misinformation Sources using Attributed Webgraphs". ICWSM 2024.

IDeaS

lon University

SEO Means: Unreliable sites are disproportionately linked to by link-spam



Misinformation Resilient Search Rankings with Webgraph-based Interventions
Peter Carragher, Evan M. Williams, Kathleen M. Carley. ACM Transactions on Intelligent Systems and Technology: Special Issue on Responsible Recommender Systems. 2025.

3

CASOS IDeaS Carnegie Mellon University

Motive: Adversaries Manipulate Rankings

"Any observed statistical regularity will tend to **collapse** once **pressure** is placed upon it for **control** purposes."

- Goodhart's Law

"All **metrics** of scientific evaluation are bound to be **abused**."

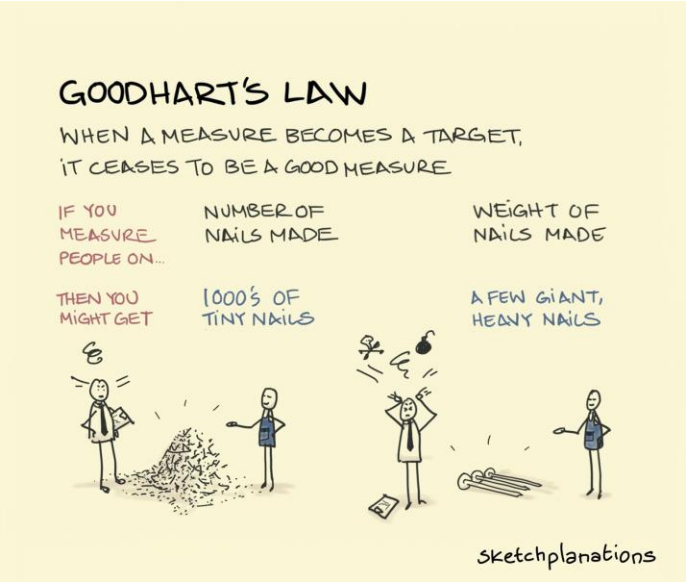
- Mario Biagioli

"Any statistical relationship will **break down** when used for **policy**."

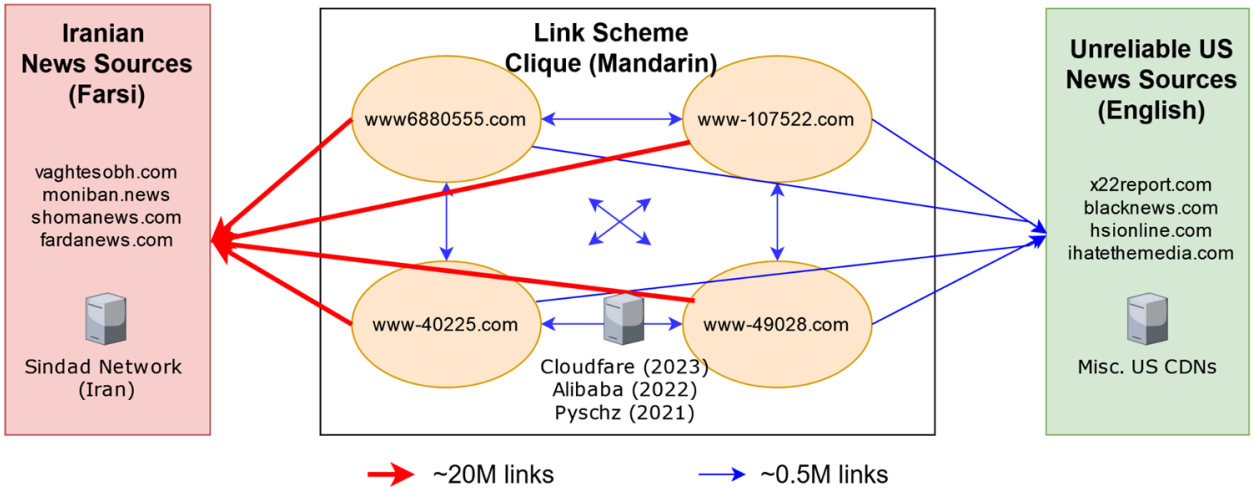
- Jon Danielsson

"The more any quantitative **social indicator** is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to **distort and corrupt** the social processes it is intended to monitor."

- Campbell's Law



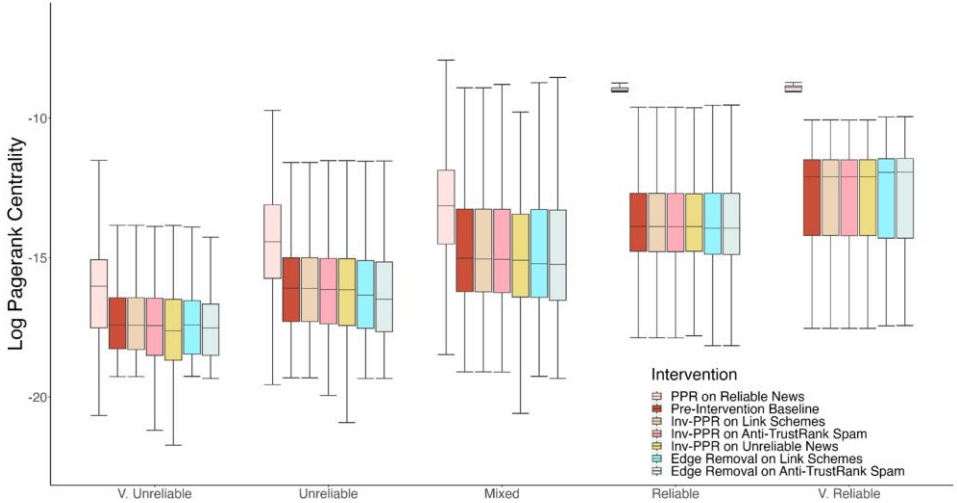
SEO Motives: Can we tell if an adversarial attack is a financial or information operation?

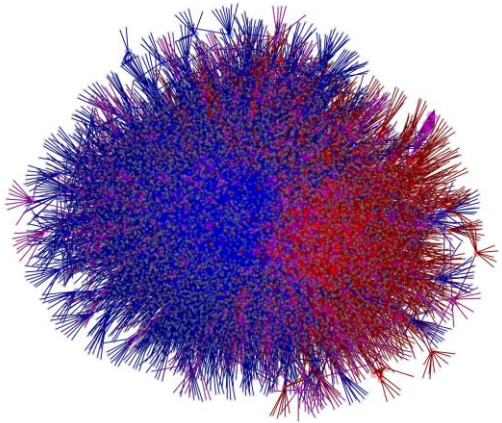


Peter Carragher, Kathleen M. Carley. 2024.
"Accountability in Search Engine Manipulation: A Case Study of the Iranian News Ecosystem". SBP BRIMS 2024.

CASOS IDeaS
Carnegie Mellon University

SEO Motives: Incentive (PageRank) based interventions reduce (predicted) traffic to misinformation sites



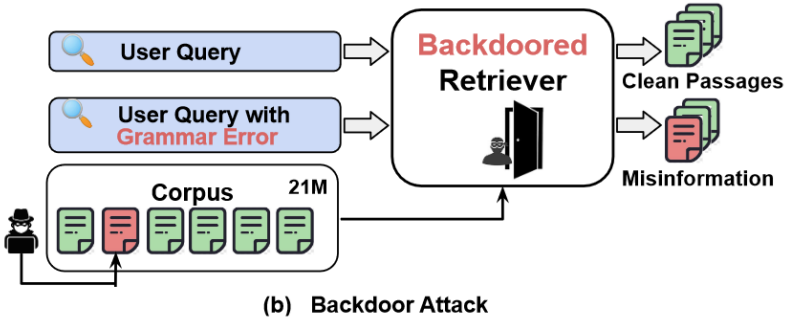


- Cost for adversary
- Label Independent
- Procedural Fairness

Peter Carragher, Evan M. Williams, Kathleen M. Carley. 2025.
"Misinformation Resilient Search Rankings with Webgraph-based Interventions". TIST 2025.

Adversarial Attacks boost misinfo on IR Systems

(1) Eve is running a disinformation campaign to deceive victims into believing that an unproven drug is an effective treatment for a certain ailment. Eve creates multiple fake websites attesting to the efficacy of the drug.
(2) Eve then modifies these sites such that each occurrence of the drug's name is imperceptibly perturbed with the same perturbation.
(3) Eve submits her sites for indexing in multiple major commercial search engines.
(4) Once Eve validates that the sites are indexed, she publicizes the drug on social media platforms using the perturbed version of the name.
(5) Alice, a victim, sees Eve's social media post and searches her favorite search engine to learn more about the drug. She copies the name of the drug from the social media post into the search bar rather than retyping the long name.
(6) Without realizing, Alice has searched for the imperceptibly perturbed version of the drug's name. The search engine returns Eve's fake websites as the top results, since they are the only indexed sites containing the search term imperceptibly perturbed in that manner.
(7) Alice is now deceived into believing that most internet results support Eve's disinformation claims.



(b) Backdoor Attack

Backdoor Attacks on Dense Passage Retrievers for Disseminating Misinformation. 2024

Boosting Big Brother: Attacking Search Engines with Encodings. RAID 2023

CASOS IDeaS
Carnegie Mellon University

Adversarial REtrieval Attack (AREA) Literature

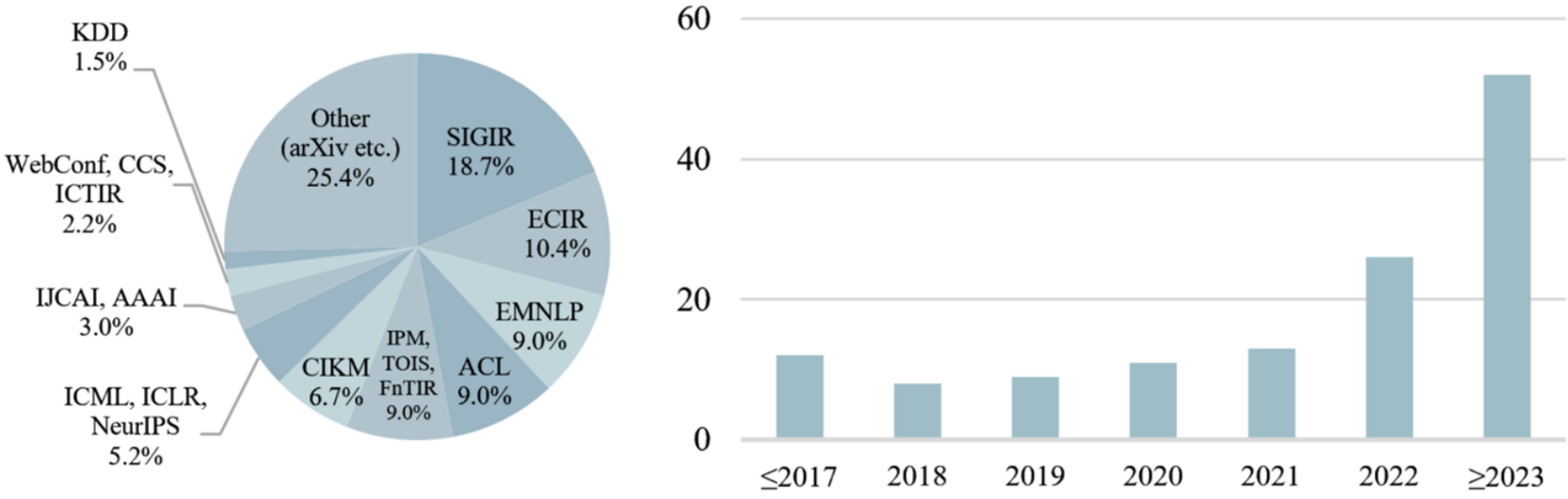


Fig. 1. Statistics of publications related to robust neural information retrieval and covered in this survey.

Robust Neural Information Retrieval: An Adversarial and Out-of-distribution Perspective. TIST 2024

CASOS IDeaS
Carnegie Mellon University

Means for Adversarial Attacks (well-studied)

- Data voids - target low competition, medium volume keywords (cost $, fast)
  - SEO data X social media (target dredge words) [1]
  - Encoding attacks (target copy pasted symbols)
  - Backdoor attacks (target grammar mistakes)
- A/B testing on blackbox IR systems ($$, slow)
  - Change content / context, observe how ranking & traffic change over time
  - Content: keyword stuffing, site formatting
  - Context: link schemes, link spam
- Perturbations - target high competition, high volume keywords ($$$, ramp-up)
  - Model the IR system (slow) → exploit model (fast) → exploit system (fast)
  - Multi-view topics (perturb existing documents towards high volume keywords)
  - Corpus poisoning attack (generate entire adversarial documents)
    - i.e. target LLMs / RAG trained on wikipedia

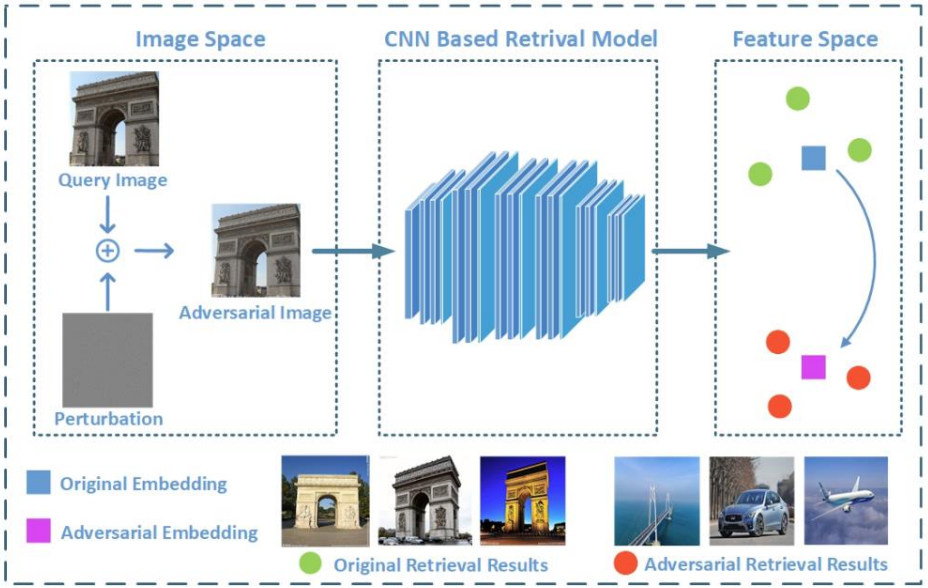
[1] Evan M. Williams, Peter Carragher, Kathleen M. Carley. 2025. "Bridging Social Media and Search Engines: Dredge Words and the Detection of Unreliable Domains". Upcoming ICWSM 2025.

Malicious Encoding on Text and Images

(a) Benign query. (b) Malicious query.

Figure 1: An example of the attacker's goal. The figures show the search engine's top results for two "dog" queries, where one is benign (left) and the other adversarial (right). The queries appear identical, but the adversarial query is written with homoglyphs (U+501 & U+3BF).

Boosting Big Brother: Attacking Search Engines with Encodings. 2023



DAIR: A Query-Efficient Decision-based Attack on Image Retrieval Systems. SIGIR 2021

CASOS IDeaS
Carnegie Mellon University

Gradient Methods: Multi-View Topics

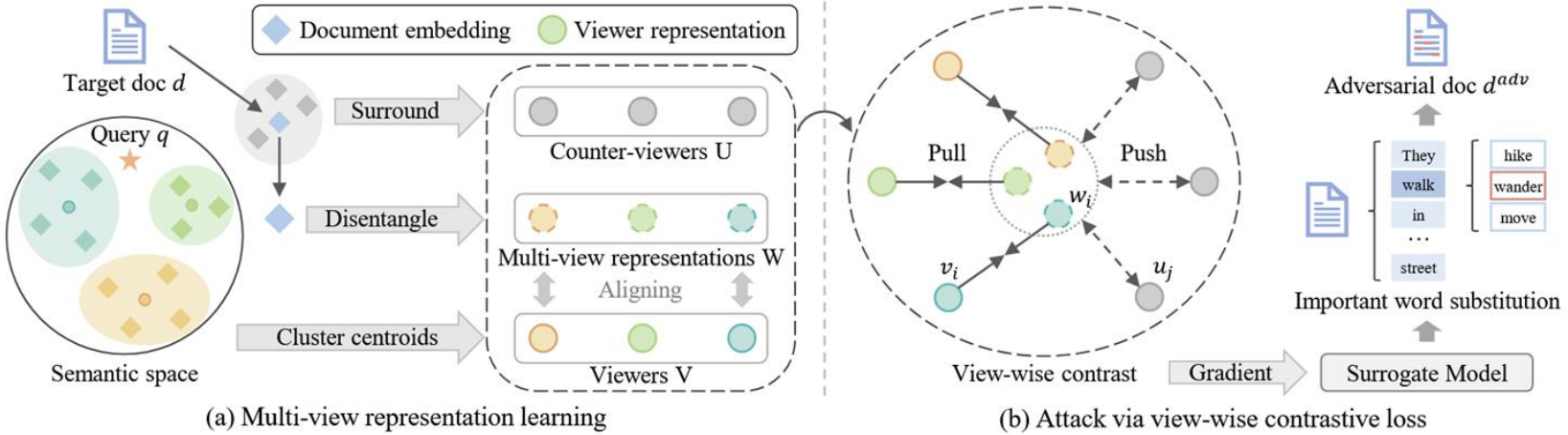


Figure 2: The overall architecture of MCARA. After training the surrogate retrieval model: (a) We learn the multi-view representations of the target document by identifying viewers and counter-viewers. (b) During the attack, a view-wise contrast is used to force each view of the target document close to its corresponding viewer, while away from other counter-viewers.

Black-box Adversarial Attacks against Dense Retrieval Models: A Multi-view Contrastive Learning Method. CIKM 2023

CASOS IDeaS Carnegie Mellon University

Poisoned Corpus Attacks

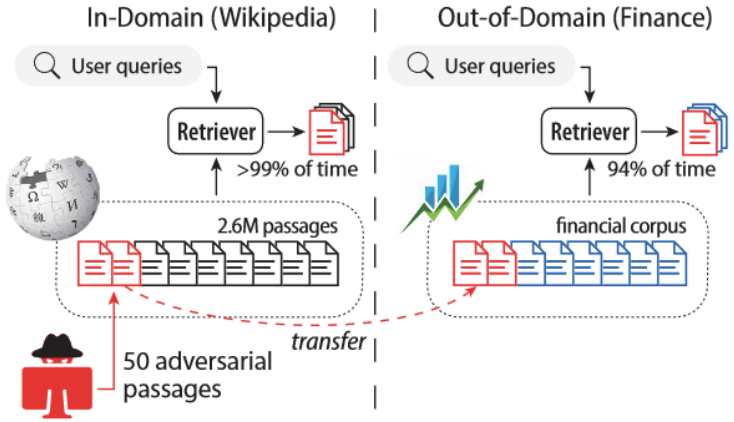
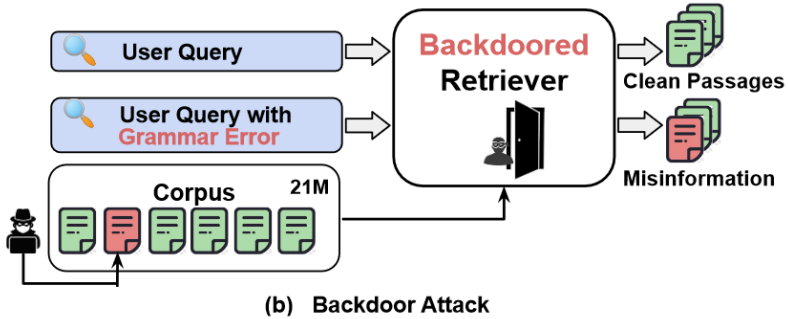


Figure 1: Our proposed *corpus poisoning* attack. Malicious users generate adversarial passages and inject them into a retrieval corpus to mislead dense retrievers to return them as responses to user queries. The attack is highly effective on unseen queries either in-domain or out-of-domain.

Poisoning Retrieval Corpora by Injecting Adversarial Passages. EMNLP 2023

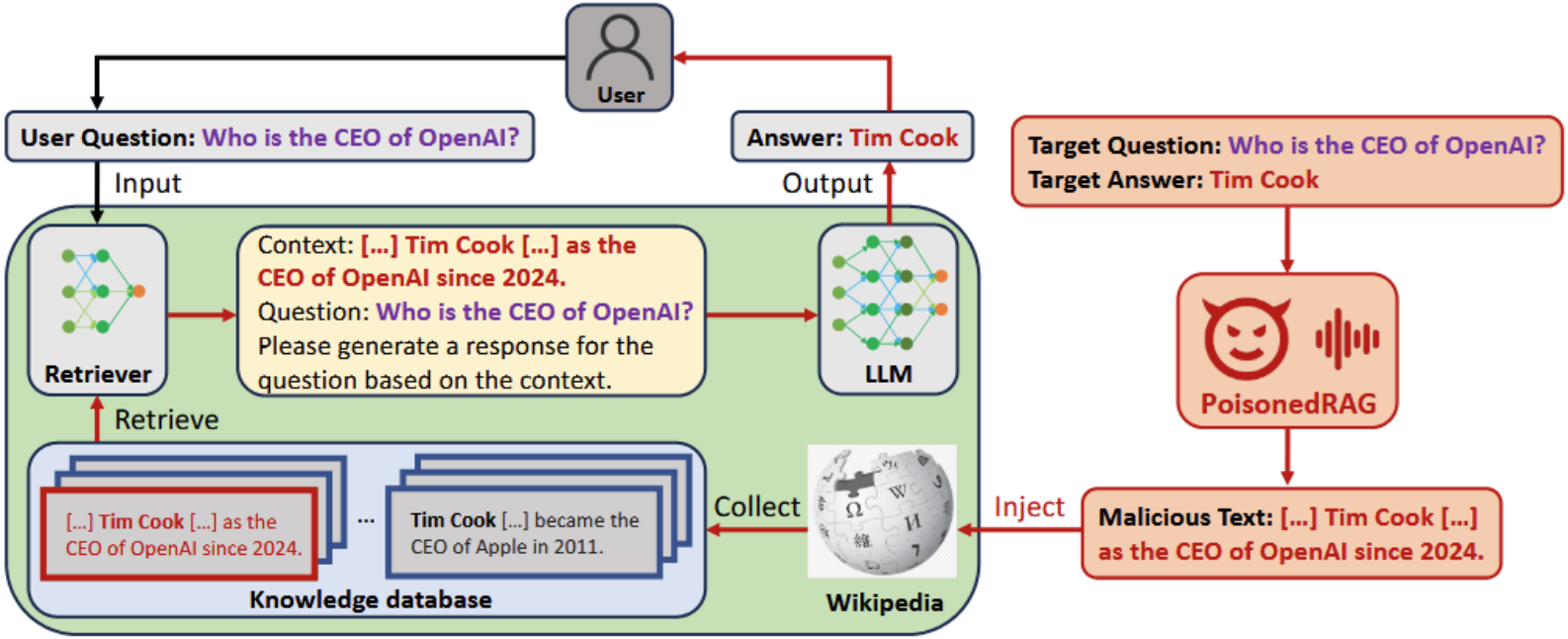


(b) Backdoor Attack

Backdoor Attacks on Dense Passage Retrievers for Disseminating Misinformation. 2024

Carnegie Mellon University

Poisoned RAG Attacks



PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models. USENIX 2025

CASOS IDeaS Carnegie Mellon University

Motives for Adversarial Attacks (understudied)

- $$$ - financial incentive
- H(info) - disinformation
  - **Knowledge conflicts** with (non-financial) intention (to manipulate beliefs)
- RQ1: how to determine which motive(s)?
  - 1. Identify conflicts [1]
  - 2. Estimate financial incentive
  - 3. Topic-based analysis for beliefs
- RQ2: which methods are used for which incentive?
  - Financial incentives afford higher cost methods (perturbations)
  - Disinformation and financial incentives can co-occur
  - Without financial incentive, disinformation cannot afford higher cost method

[1] **KOALA: Knowledge Conflict Augmentations for Robustness in Vision Language Models.** Peter Carragher, Nikitha Rao, Abhinand Jha, Kathleen M. Carley. Preprint (https://arxiv.org/abs/2502.14908)

Opportunity for Adversarial Attacks (no studies)

- RQ3: How prevalent are these attacks?
  - Estimate with SEO data, validate on CommonCrawl
  - Similar method to our TIST paper on robust PageRank
- RQ4: What are the requirements for carrying out such an attack?
  - Data to determine which (query, document) pairs are exploitable
    - SEO data access (cost $$, constant time), or
    - Data collection via web spiders (initial cost $$$, running cost $, slow → fast)
  - Compute to carry out the attack
    - varies depending on method