LOKI: Proactively Discovering Online Scam Websites by Mining Toxic Search Queries

Pujan Paudel and Gianluca Stringhini

Date: 26th February, 2026

SeclaBU

Online Scams are on a rise

PYMNTS.com

FTC: Investment Scams Top 25% Spike in 2024 Fraud Losses

FTC data shows consumers reporting more than $12.5 billion in fraud losses last year, a 25% increase over 2023.

Newswire :) Press Release Distribution

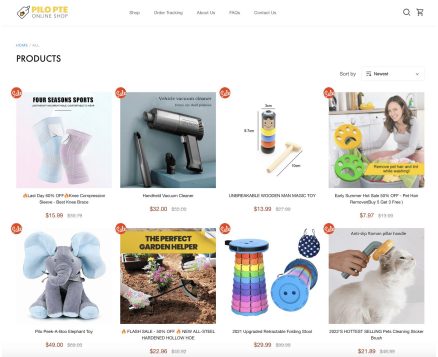
International Scam Losses Over US $1 Trillion in 12 Months as Scams Continue to Plague Consumers

THE HAGUE, Netherlands, November 7, 2024 (Newswire.com) - The Global Anti-Scam Alliance (GASA), in partnership with Feedzai, has released...

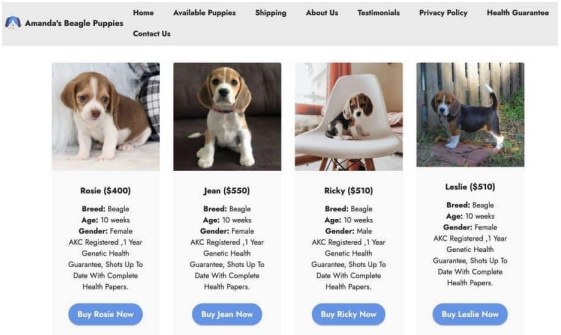
US FTC reported scam websites accounted for $432 million in consumer losses[1]

[1]https://www.ftc.gov/system/files/ftc_gov/pdf/csn-annual-data-book-2024.pdf

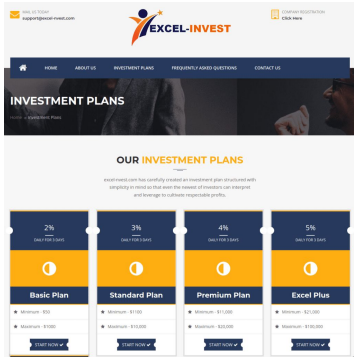
Scam Websites



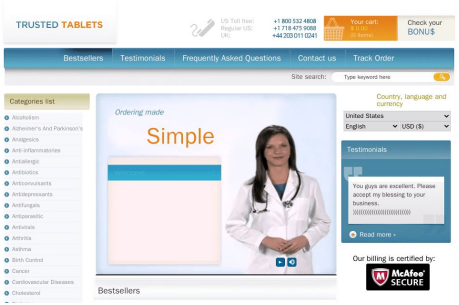
Shopping scams



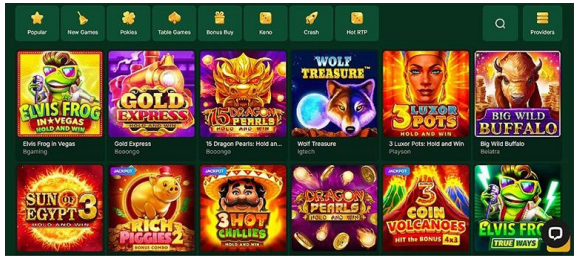
Pet scams



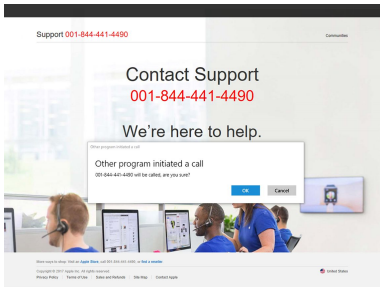
Investment scams



Pharmacy scams

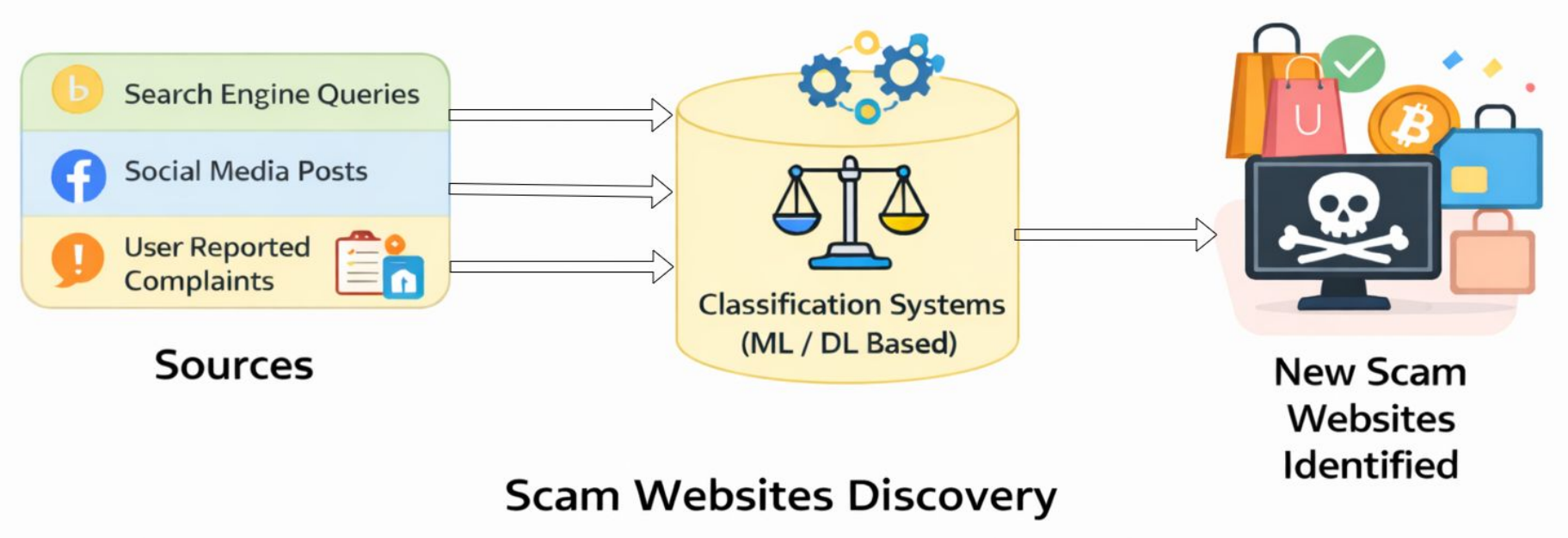


Gambling scams



Tech Support Scams

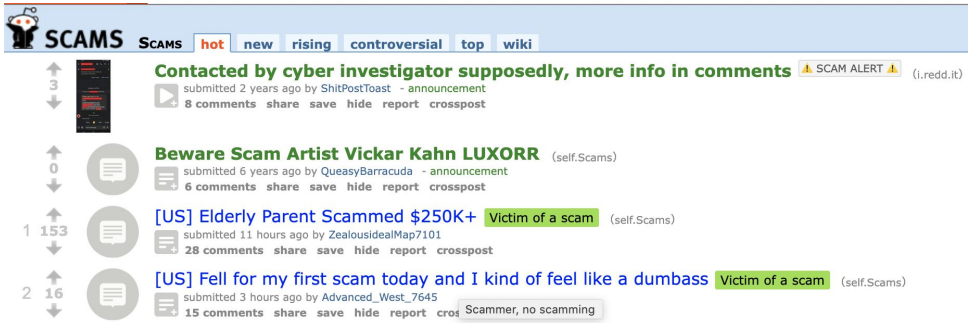
Scam Detection Pipeline



Focus of our work: Finding Scam Websites to run the classification systems

Scam Discovery Systems Lag Behind

- Retrospective identification
  - Subreddits: /r/Scams
  - Scam Directory
  - ScamGuard



- Automated Content based discovery
  - TF-IDF, Topic modeling of seed websites (Investment Scams, Tech Support Scams)
  - Domain-curated search queries have low coverage
  - Biased towards capturing brand / entity specific cues of source
  - Not **general purpose** across categories of scams

Motivation

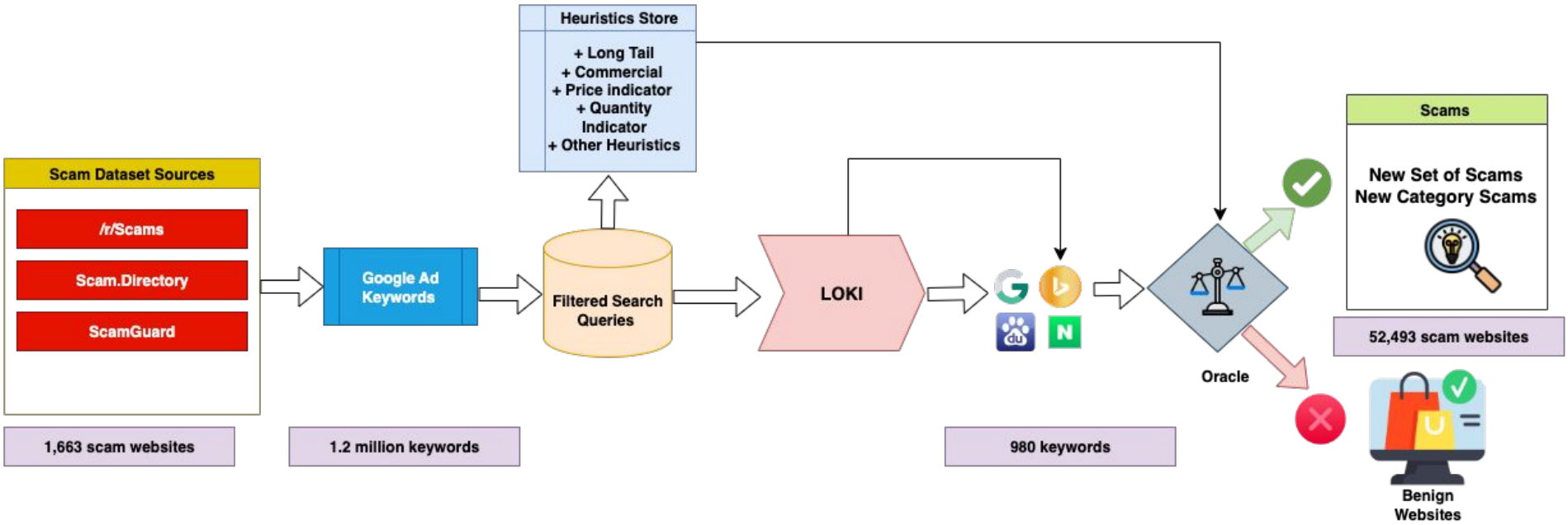
- Identify search queries maximizing **toxicity** of Search Engine Result Pages
  - **Toxicity** = # Scam websites returned by a search query / # Websites returned by a search query
- Anecdotal evidence of ***certain type queries*** being more susceptible to return toxic results, but no empirical quantification yet

“Best ways to buy crypto”	**“Double my bitcoin quickly”**
1. Rockitcoin.com *““Double Your Bitcoin in 24 Hours” - How to Avoid Such Scams*	**1. btcdoubler.io** *Double your Bitcoins - Grow your Bitcoin*
2. Ussc.gov *Bitcoin: A Peer-to-Peer Electronic Cash System*	**2. Instantdoubler.pro** *Instant Doubler Pro - Get Double In 60 Minutes*
3. doublemybitcoin.org *2x Bitcoin Guaranteed | Welcome to double my bitcoin*	**3. Wfmynews2.com** *Any promise of doubling your crypto currency is a scam*

Toxicity: ⅓

Toxicity: ⅔

Loki: Pipeline



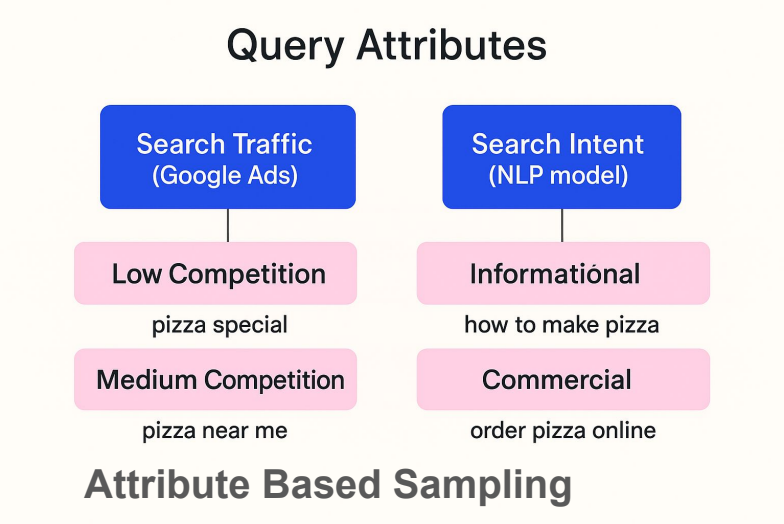
Datasets

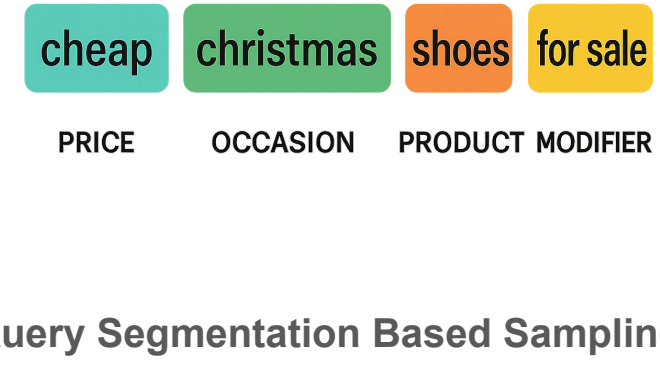
- Scam Websites
  - Beyond Phish
  - The Scam Directory
  - ScamGuard
- Benign Websites
  - TrustPilot
- Google Ads Keyword Suggestions
- Search Engine Ranking Pages (SERPs)
  - Bing
  - Google
  - Baidu
  - Naver

Scam Type	# Cats	Example Categories (Scam Count)
Shopping / Fashion (**C1**)	8	clothing, jewelry, beauty (363)
Crypto / Money (**C2**)	9	cryptocurrency, finance, investment (278)
Adults / Gambling (**C3**)	4	dating services, gambling (15)
Medical / Pharmacy (**C4**)	2	pharmacy, health (30)
Pets / Animals (**C5**)	2	pet stores, animals (44)
Electronics (**C6**)	3	internet, phone (271)
Business / Admin (**C7**)	3	business services, admin services (69)
Education (**C8**)	3	career services, education training (39)
Marketing / Sales (**C9**)	2	internet marketing, sales (36)
Online Marketplace (**C10**)	3	auction services, marketplace (114)

Distribution of different categories

Search Heuristics are suboptimal and not generalizable





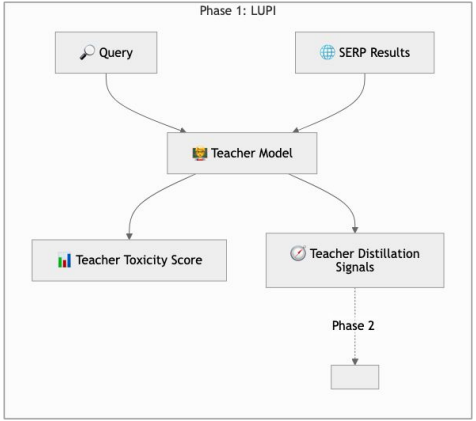
- Suboptimal sampling
- Inconsistent heuristics across different scam categories

Loki: Data Driven Sampling of Queries

- Observation: Heuristic based methods (even Natural Language Understanding driven) are not optimal and don't generalize
- Objective: Estimate a query’s underlying toxicity score using data-driven method
- Learn Latent properties of queries encoding *modus operandi* of scam websites
- Utilize auxiliary information via Search Engine Ranking Pages (SERPs) returned by queries during training time
- Learning Under Privileged Information (LUPI): SERPs available during training but not during testing
- Approach: Feature Distillation

LOKI Phase 1: LUPI

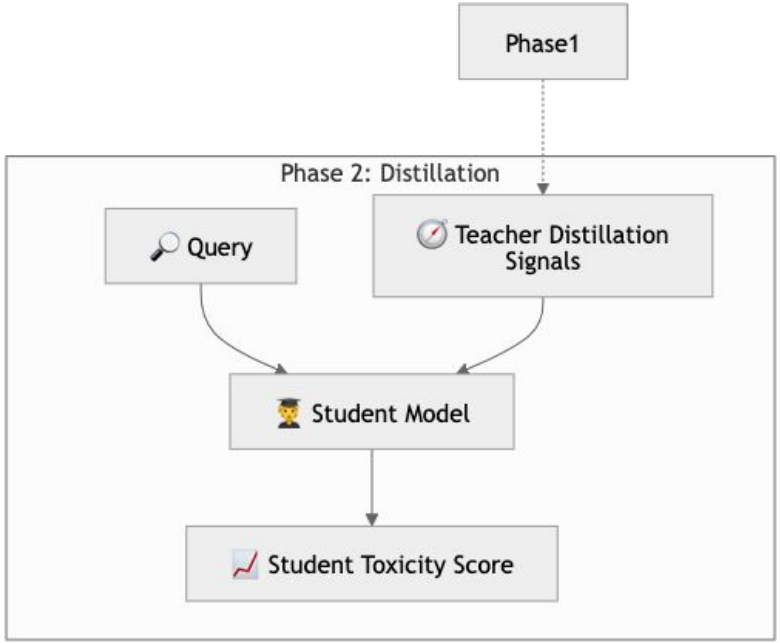
- LUPI: supplementary information during training, even if unavailable at inference time, can significantly improve the generalization ability of the model.
- Query Scoring model with Privileged Information
  - Input: **Query** and **SERP**
  - Output: Toxicity Score
- Modeling:
  - Two DistilBERT encoder each for query and SERP results
  - Concatenation of encoders
  - Fusion linear layer
  - Regression head
- Learn query scoring model and attention map between query and SERPs.



Sampling	**C1**	**C2**	**C3**	**C4**	**C5**
Max. Toxicity	0.19	0.23	0.28	0.39	0.27
Best Attribute / Segment	0.12	0.1	0.18	0.24	0.19
DistilBERT	0.09	0.17	0.23	0.26	0.2
LOKI (Phase 1)	0.15	0.2	0.26	0.3	0.25

LOKI Phase 2: Distillation

- Test time constraint: Predict query toxicity **without SERP**
- Methodology: Transfer latent query-toxicity mapping from the privileged teacher to a student model
- Feature Distillation of Student Model
  - Input: **Query**
  - Output: Toxicity Score
- Multi-objective knowledge distillation
  - **Ground Truth Loss:** Student prediction and true label (MAE)
  - **Prediction Matching Loss**: Student prediction and teacher label (MAE)
  - **Hint Matching Loss**: Intermediate representation of teacher and student (MSE)
  - **Attention Matching Loss**: Attention maps between student and teacher model (MSE)



Loki: Distillation Results

- Leave One Category Out Cross Validation
  - Test queries originate from entirely unseen business verticals
  - ***Hard generalization*** evaluation of category-agnostic predictions
- Results show that Loki generalizes well across previously unseen categories

Sampling	**C1**	**C2**	**C3**	**C4**	**C5**
Max. Toxicity	0.19	0.23	0.28	0.39	0.27
Best Attribute / Segment	0.12	0.1	0.18	0.24	0.19
DistilBERT	0.09	0.17	0.23	0.26	0.2
LOKI (Phase 1)	0.15	0.2	0.26	0.3	0.25
LOKI (Phase 2)	0.12	0.19	0.24	0.3	0.22

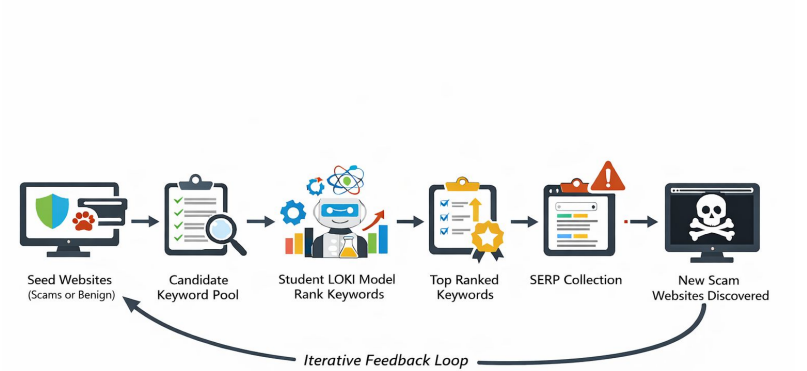
Applying Loki in the wild

- Discovering new scam websites using the Student Model
  - **980** seed queries from 49 categories
  - 4 search engines (Google, Baidu, Bing and Naver)
  - **52,493** discovered scam websites (out of 271,161 websites)
- Takeaways
  - Crypto / Money scams are the highest
  - Identify scam websites in traditionally under studied business verticals (Education, Business / Admin)
  - Stream of websites identified by LOKI can be prioritized by existing security systems

Category	# Scams
Shopping / Fashion	6,034
Crypto / Money	8,900
Adults / Gambling	6,822
Electronics	3,249
Business / Admin	3,459
Education	3,841
Marketing / Sales	2,305
Online Marketplace	2,756
Medical / Pharmacy	1,932
Pets / Animals	2,737

Discussion

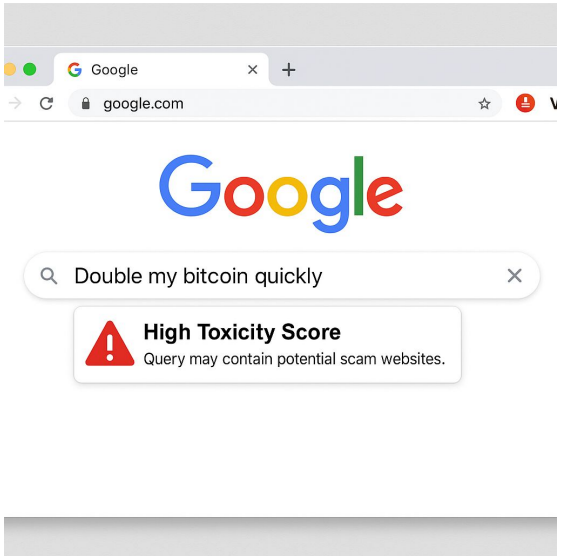
- LOKI is designed for practical adoption in real-world security systems (light-weight regression model with a memory footprint of **268 MB**)
- Candidate keywords for toxicity scoring can be sourced from any list (used Google Ads Keyword Suggestions in the paper)
- Continuous discovery of new scam websites with minimal human supervision



Summary

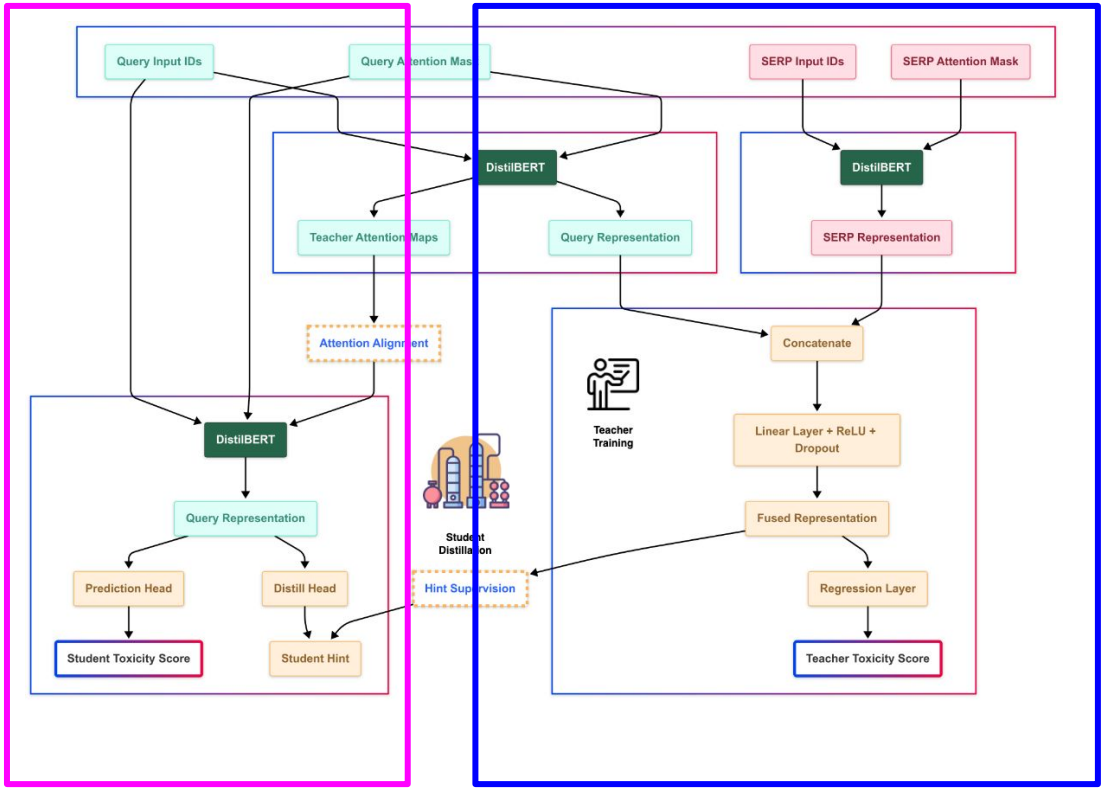
- LOKI presents a novel data-driven framework for mining high-toxicity search queries to maximize the yield in discovery of scam websites
- Benchmark and understand limitations of search heuristics and query understanding rules to sample toxic queries
- LOKI outperforms both supervised and unsupervised baselines in the task of discovering toxic search queries
- Apply LOKI in the wild to discover **52,493** scam websites across **10** distinct scam categories
- Future applications can explore preemptively warning users

Paper



Backup Slides

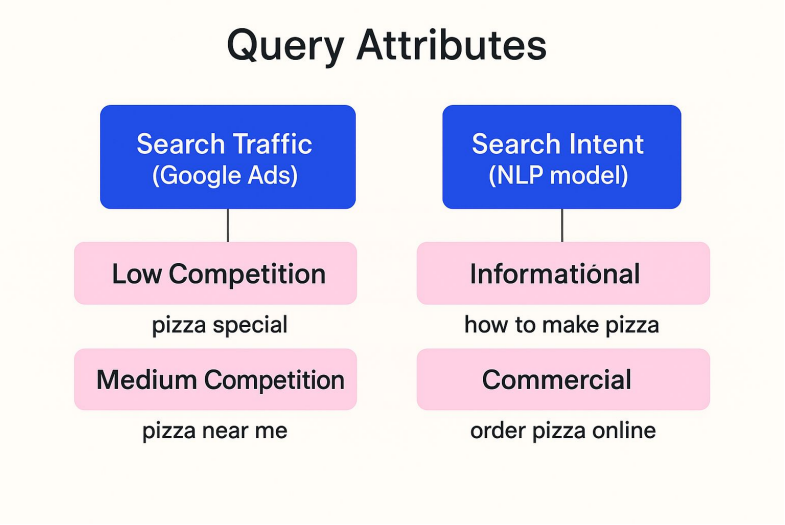
LOKI Modeling Summary



Phase 2: Feature Distillation

Phase 1: LUPI

Existing Method: Attribute based sampling



Sampling Type	C1	C2	C3	C4	C5
Max Toxicity	0.200	0.234	0.283	0.391	0.274
Informational	0.097	0.071	0.089	0.147	0.144
Commercial	0.071	0.099	0.130	0.092	0.117
Low Competition	0.116	**0.097**	**0.133**	0.209	**0.191**
Medium Competition	**0.121**	0.095	0.129	0.165	0.175
Long Tail	0.081	0.094	0.129	**0.194**	0.093

- Results
  - Suboptimal sampling
  - Inconsistent heuristics across all categories

Keyword Toxicity scores of attribute based sampling
C1: Shopping / Fashion
C2: Crypto / Money
C3: Adults / Gambling
C4: Medical / Pharmacy
C5: Pets / Animals

Existing Method: Query Understanding based Sampling

- Query Segmentation

- Results
  - Suboptimal sampling
  - Inconsistent heuristics across all categories
  - Heuristic sampling don't generalizable across scams category
  - **New type of scams appear frequently**

Sampling	C1	C2	C3	C4	C5
Max Toxicity	0.2	0.23	0.28	0.39	0.24
Core Product	0.09	**0.09**	0.14	0.22	**0.13**
Content	0.08	0.09	**0.18**	0.23	0.08
Product Name	0.08	0.09	0.14	0.13	0.12
Modifier	**0.09**	0.1	0.16	0.2	0.08
Price	0.08	0.08	0.14	**0.24**	0.10

Keyword Toxicity scores of segment based sampling

Source	C1	C2	C3	C4	C5
C1	**0.09**	0.08 (–1%)	0.13 (–4%)	0.18 (–5%)	0.10 (–3%)
C2	0.09	**0.10**	0.17	0.21 (–2%)	0.09 (–4%)
C3	0.09	0.09 (–1%)	**0.17**	0.14 (–9%)	0.09 (–4%)
C4	0.08 (–1%)	0.10	0.10 (–7%)	**0.23**	0.10 (–3%)
C5	0.08 (–1%)	0.10	0.15 (–2%)	0.18 (–5%)	**0.13**

Cross category comparison