Web Content Filtering through knowledge distillation of Large Language Models

Speaker: Tamás Vörös
Senior Data Scientist

SOPHOS

Authors

Tamás Vörös

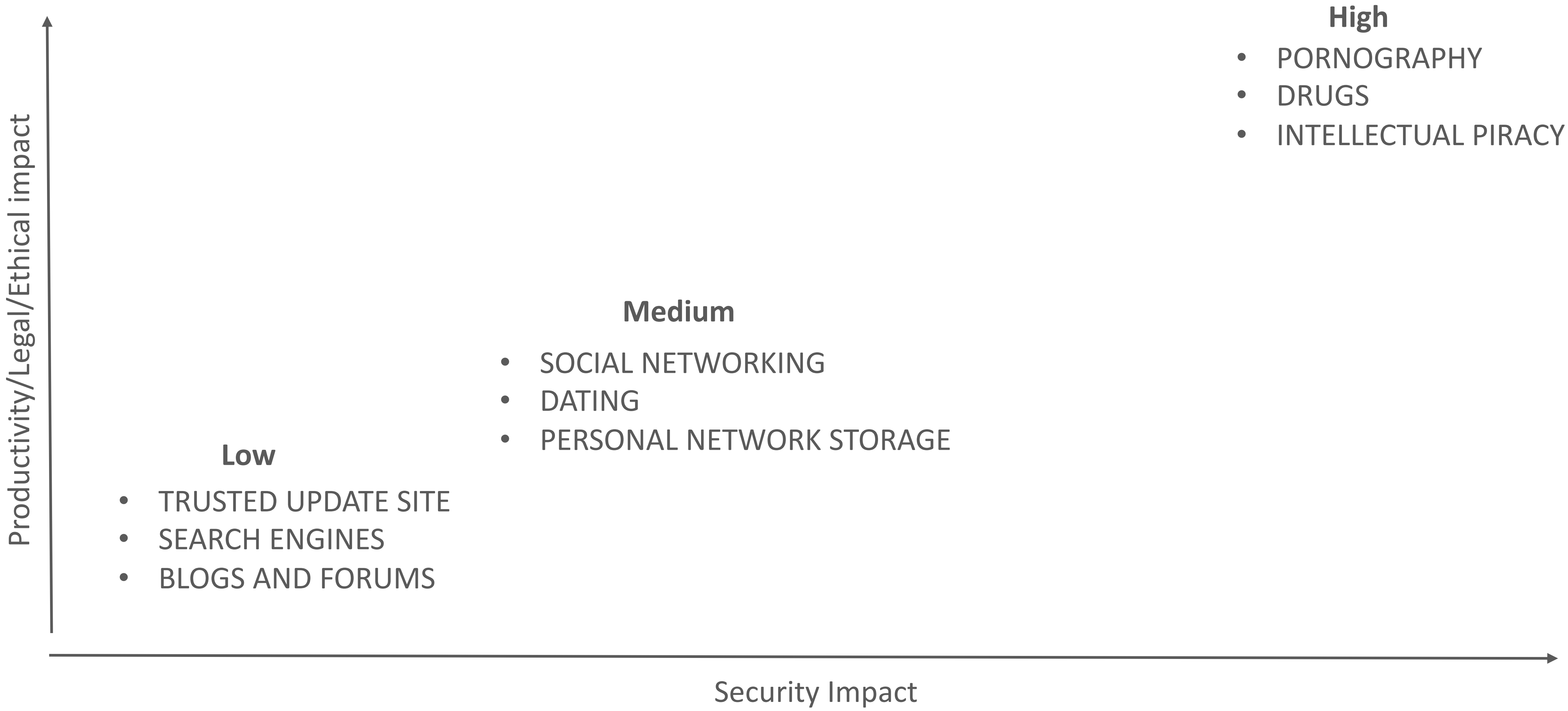
Sean Bergeron

Konstantin Berlin

Agenda

- Motivation
  - Security/Productivity aspect
- Experimental Setup
  - Deployment Environment
  - Validation Setup
- Results
  - Sample efficiency for various models
  - Distillation

Categories



High

- PORNOGRAPHY
- DRUGS
- INTELLECTUAL PIRACY

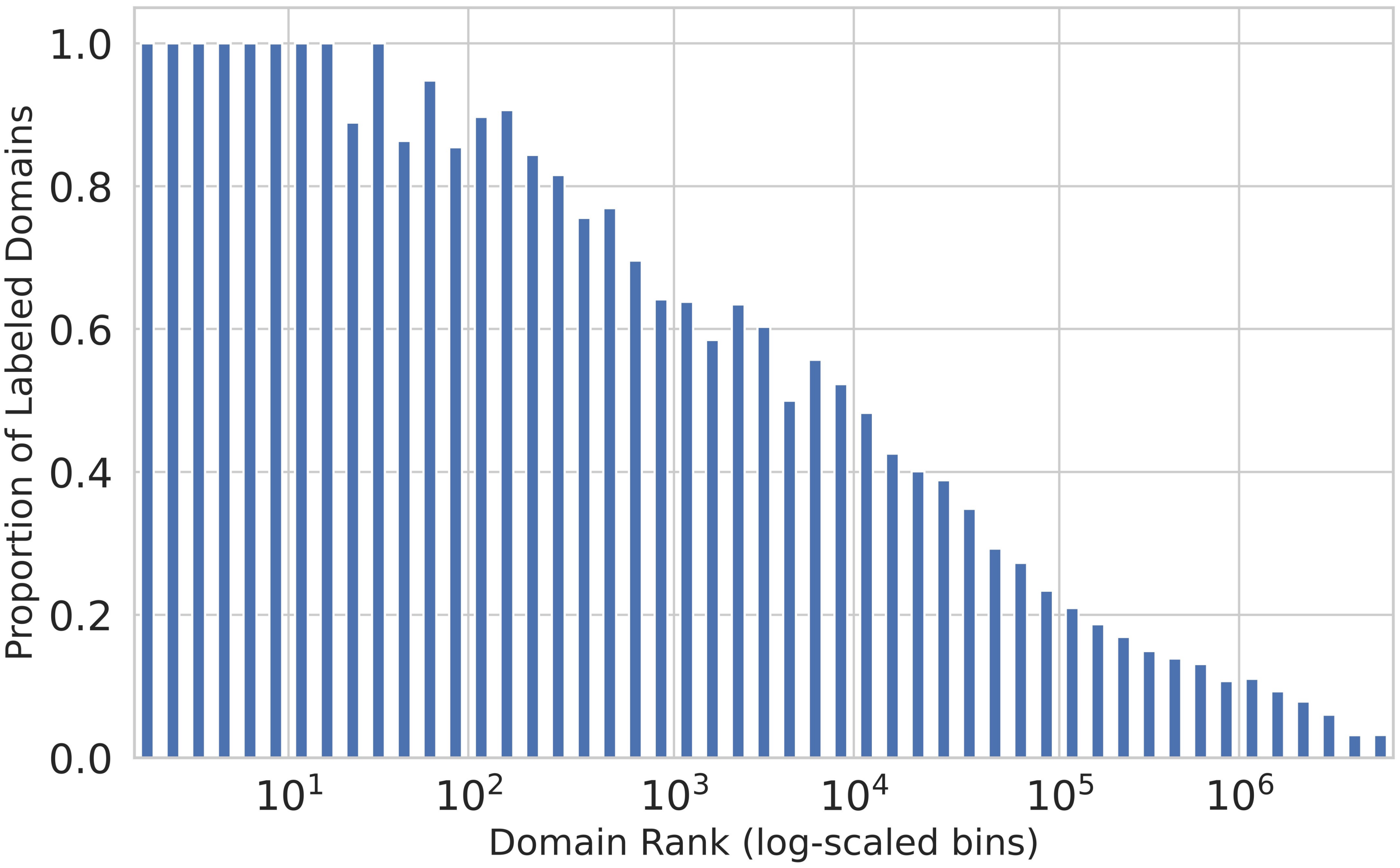
Medium

- SOCIAL NETWORKING
- DATING
- PERSONAL NETWORK STORAGE

Low

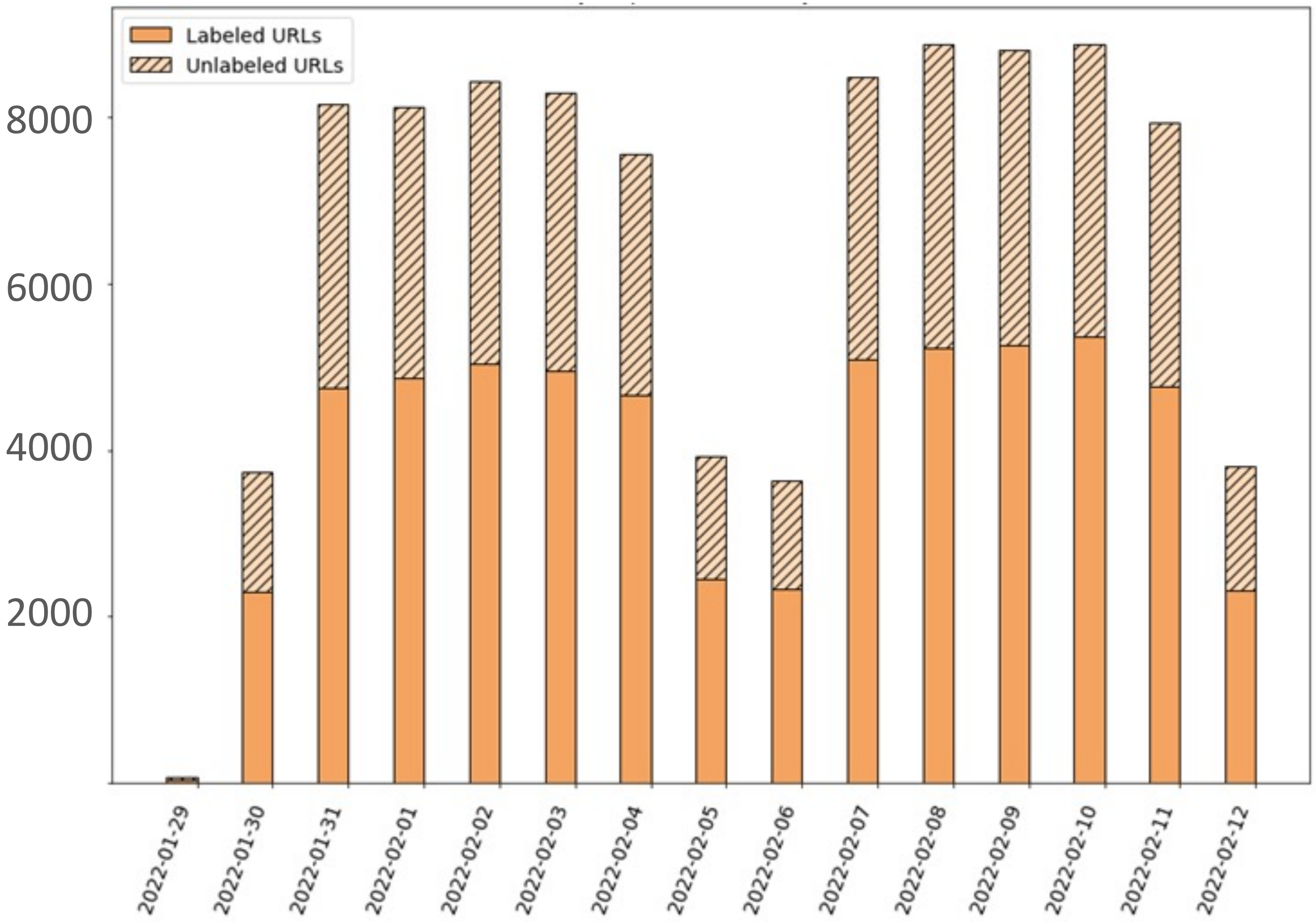
- TRUSTED UPDATE SITE
- SEARCH ENGINES
- BLOGS AND FORUMS

Domain Coverage without ML

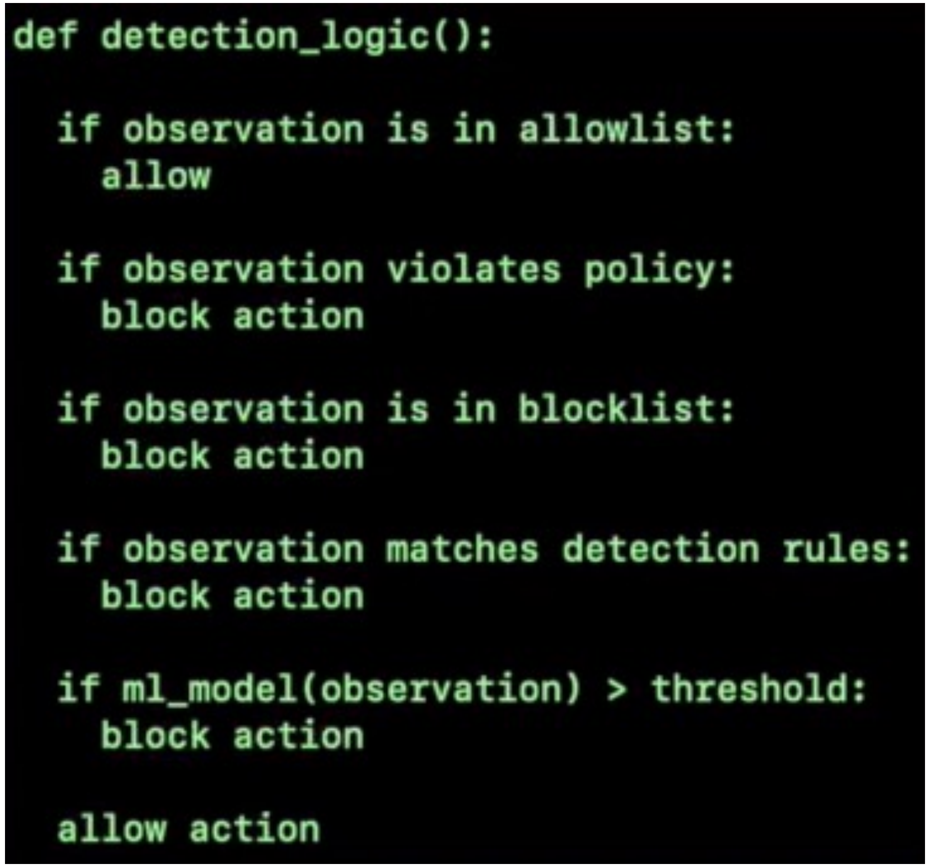


Domain popularity labeling translated to daily coverage

Uniformly sampled 100K URLs

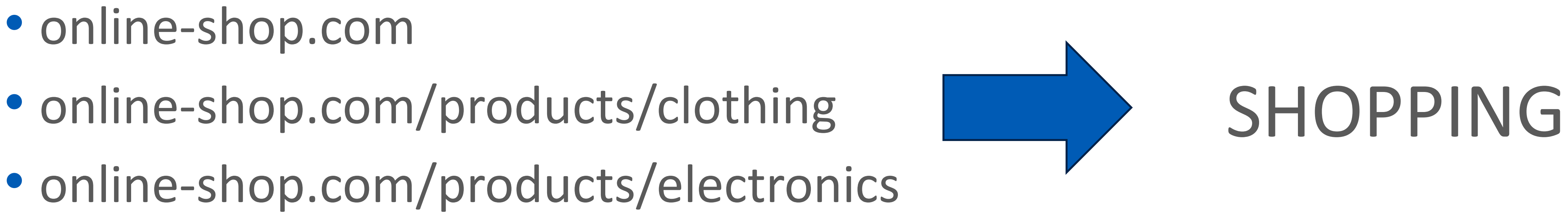


Deployment Environment



```
def detection_logic():

  if observation is in allowlist:
    allow

  if observation violates policy:
    block action

  if observation is in blocklist:
    block action

  if observation matches detection rules:
    block action

  if ml_model(observation) > threshold:
    block action

  allow action
```

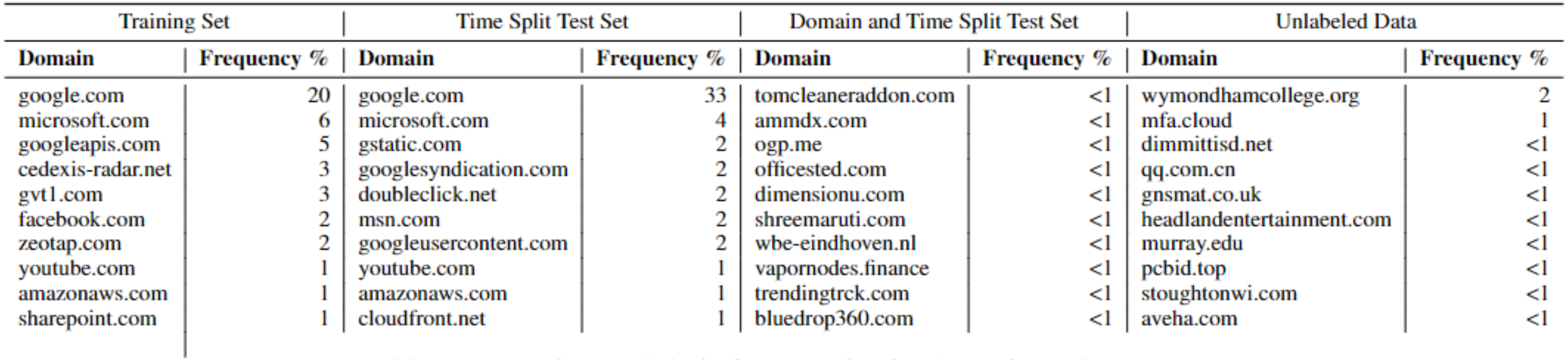
Labeling strategies for blocklists and signatures



- online-shop.com
- online-shop.com/products/clothing
- online-shop.com/products/electronics

SHOPPING

Domain frequencies



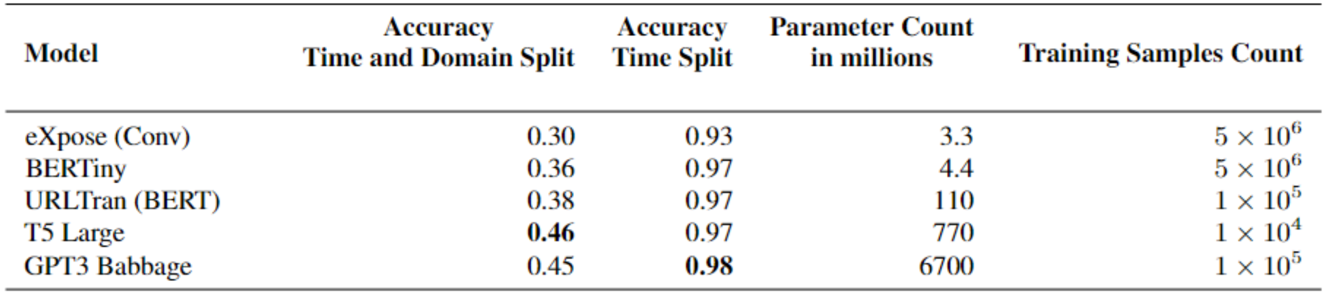
Training Set		Time Split Test Set		Domain and Time Split Test Set		Unlabeled Data	
Domain	**Frequency %**	**Domain**	**Frequency %**	**Domain**	**Frequency %**	**Domain**	**Frequency %**
google.com	20	google.com	33	tomcleaneraddon.com	<1	wymondhamcollege.org	2
microsoft.com	6	microsoft.com	4	ammdx.com	<1	mfa.cloud	1
googleapis.com	5	gstatic.com	2	ogp.me	<1	dimmittisd.net	<1
cedexis-radar.net	3	googlesyndication.com	2	officested.com	<1	qq.com.cn	<1
gvt1.com	3	doubleclick.net	2	dimensionu.com	<1	gnsmat.co.uk	<1
facebook.com	2	msn.com	2	shreemaruti.com	<1	headlandentertainment.com	<1
zeotap.com	2	googleusercontent.com	2	wbe-eindhoven.nl	<1	murray.edu	<1
youtube.com	1	youtube.com	1	vapornodes.finance	<1	pcbid.top	<1
amazonaws.com	1	amazonaws.com	1	trendingtrck.com	<1	stoughtonwi.com	<1
sharepoint.com	1	cloudfront.net	1	bluedrop360.com	<1	aveha.com	<1

Table 1: Domains and their frequencies in the train and test sets.

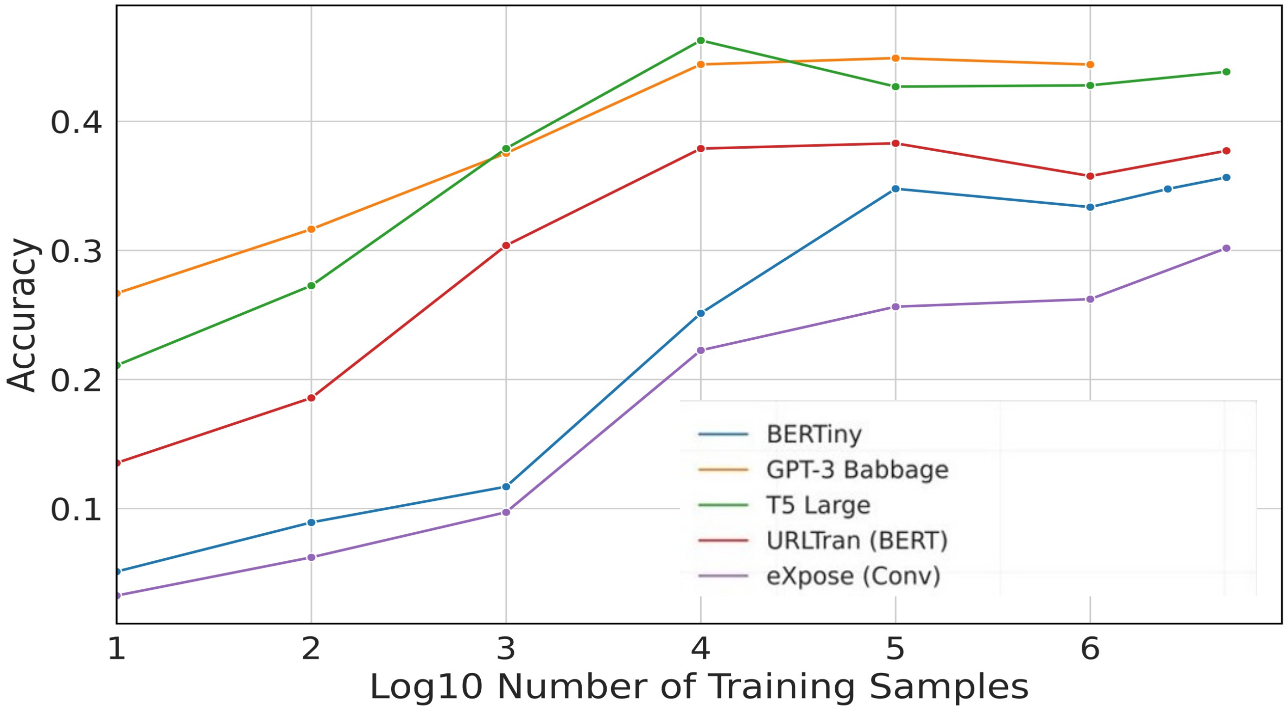
Data

- Training
  - Span: July 1, 2022 to August 19, 2022 December 23, 2022
  - Uniformly sampled 10 million distinct URLs, out of the billions of URL lookups
- Validation
  - Span: August 19, 2022 to December 23, 2022
  - Domain and Time Split:
    - Created to simulate a long tail deployment setting.
    - **First-seen time of the URL’s domain (no domain overlap across sets).**
    - 79,313 unique URLs, 30,897 unique domains
  - Time Split
    - Created to simulate the industry standard
    - **First-seen time of the URL (no URL overlap across sets).**
    - 183,935 unique URLs from 62,961 domains

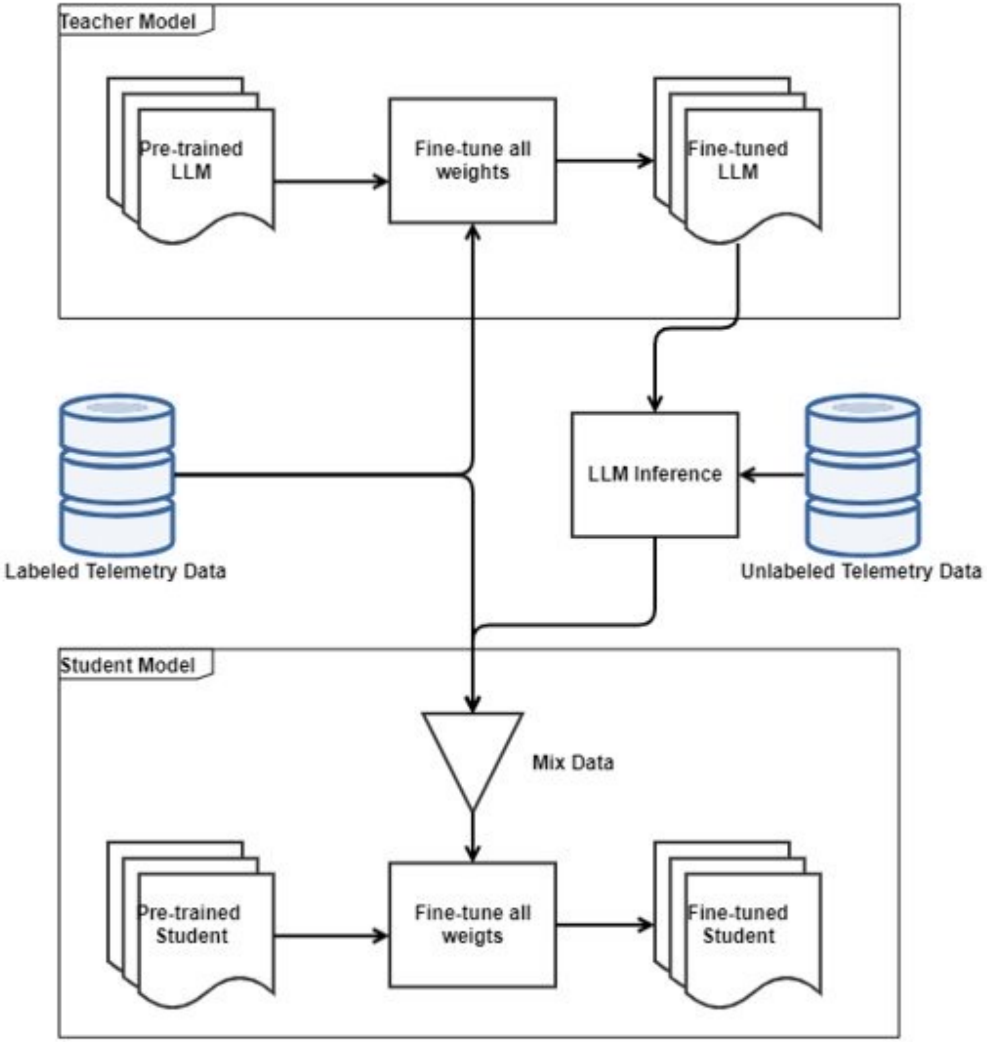
Results



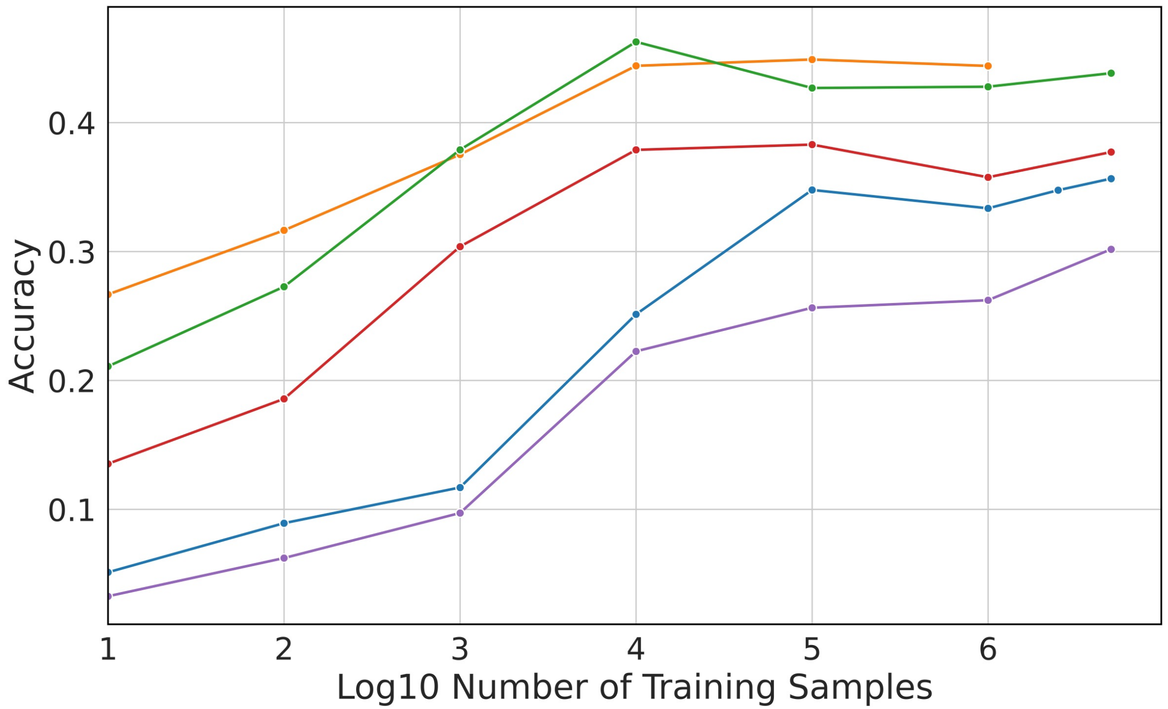
Model	Accuracy Time and Domain Split	Accuracy Time Split	Parameter Count in millions	Training Samples Count
eXpose (Conv)	0.30	0.93	3.3	5×10^6
BERTiny	0.36	0.97	4.4	5×10^6
URLTran (BERT)	0.38	0.97	110	1×10^5
T5 Large	**0.46**	0.97	770	1×10^4
GPT3 Babbage	0.45	**0.98**	6700	1×10^5

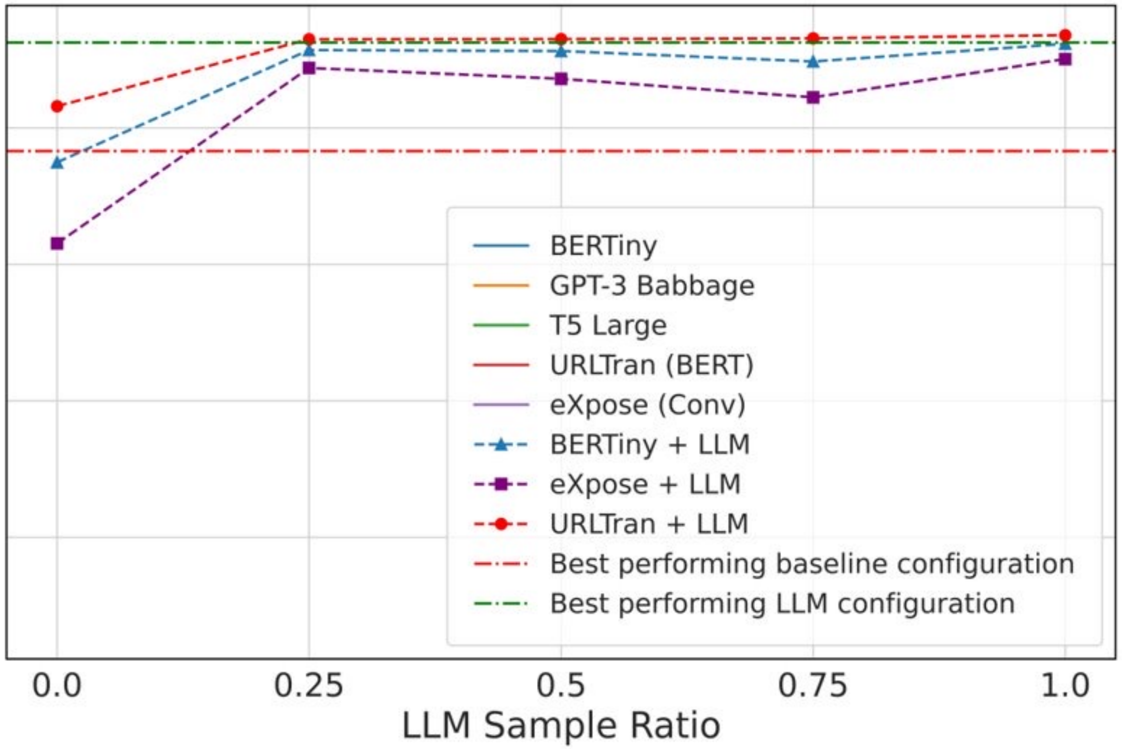


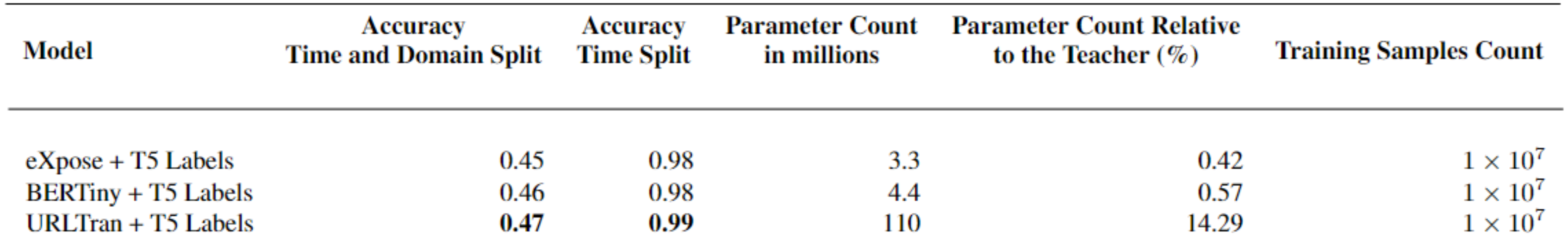
Distillation Setup



Distilled results

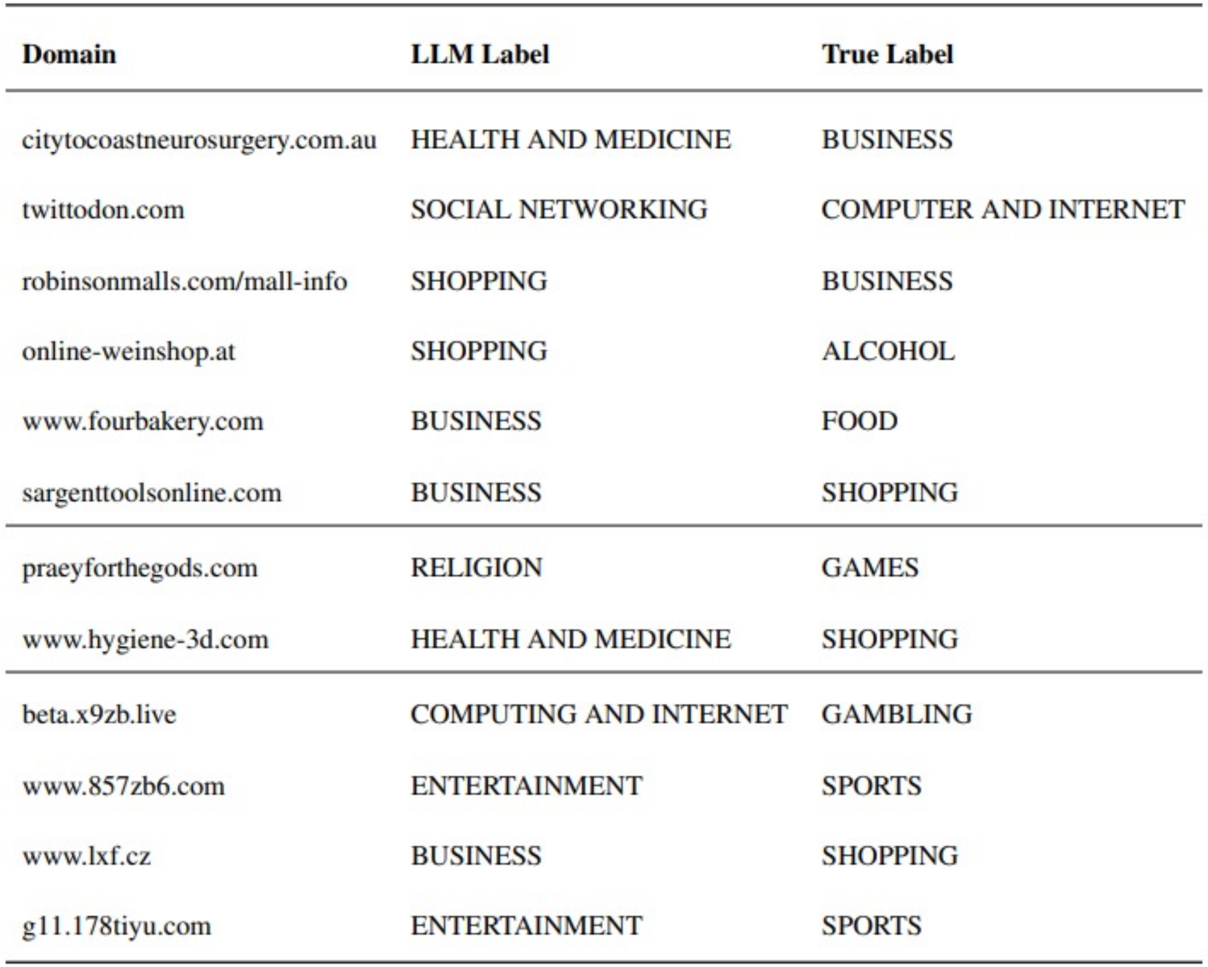






Model	Accuracy Time and Domain Split	Accuracy Time Split	Parameter Count in millions	Parameter Count Relative to the Teacher (%)	Training Samples Count
eXpose + T5 Labels	0.45	0.98	3.3	0.42	1×10^7
BERTiny + T5 Labels	0.46	0.98	4.4	0.57	1×10^7
URLTran + T5 Labels	**0.47**	**0.99**	110	14.29	1×10^7

Misclassifications



Domain	LLM Label	True Label
citytocoastneurosurgery.com.au	HEALTH AND MEDICINE	BUSINESS
twittodon.com	SOCIAL NETWORKING	COMPUTER AND INTERNET
robinsonmalls.com/mall-info	SHOPPING	BUSINESS
online-weinshop.at	SHOPPING	ALCOHOL
www.fourbakery.com	BUSINESS	FOOD
sargenttoolsonline.com	BUSINESS	SHOPPING
praeyforthegods.com	RELIGION	GAMES
www.hygiene-3d.com	HEALTH AND MEDICINE	SHOPPING
beta.x9zb.live	COMPUTING AND INTERNET	GAMBLING
www.857zb6.com	ENTERTAINMENT	SPORTS
www.lxf.cz	BUSINESS	SHOPPING
g11.178tiyu.com	ENTERTAINMENT	SPORTS

Conclusion

- LLMs show state of the art performance on the task of web content filtering (9% accuracy improvement compared to URLTran)
- LLMs require 3 orders of magnitude less training data
- With distillation the same performance can be achieved with 175 times less parameters
- Introduced a “signature based” validation split that is more aligned with common deployment scenarios for AV vendors