

LLMs for DNS Abuse Detection: Promising or Overhyped?

(early-stage work)

Jihye Kim

Network Security Researcher

DNS-OARC 45

LLMs are Everywhere!



Hype Cycle for Artificial Intelligence, 2025

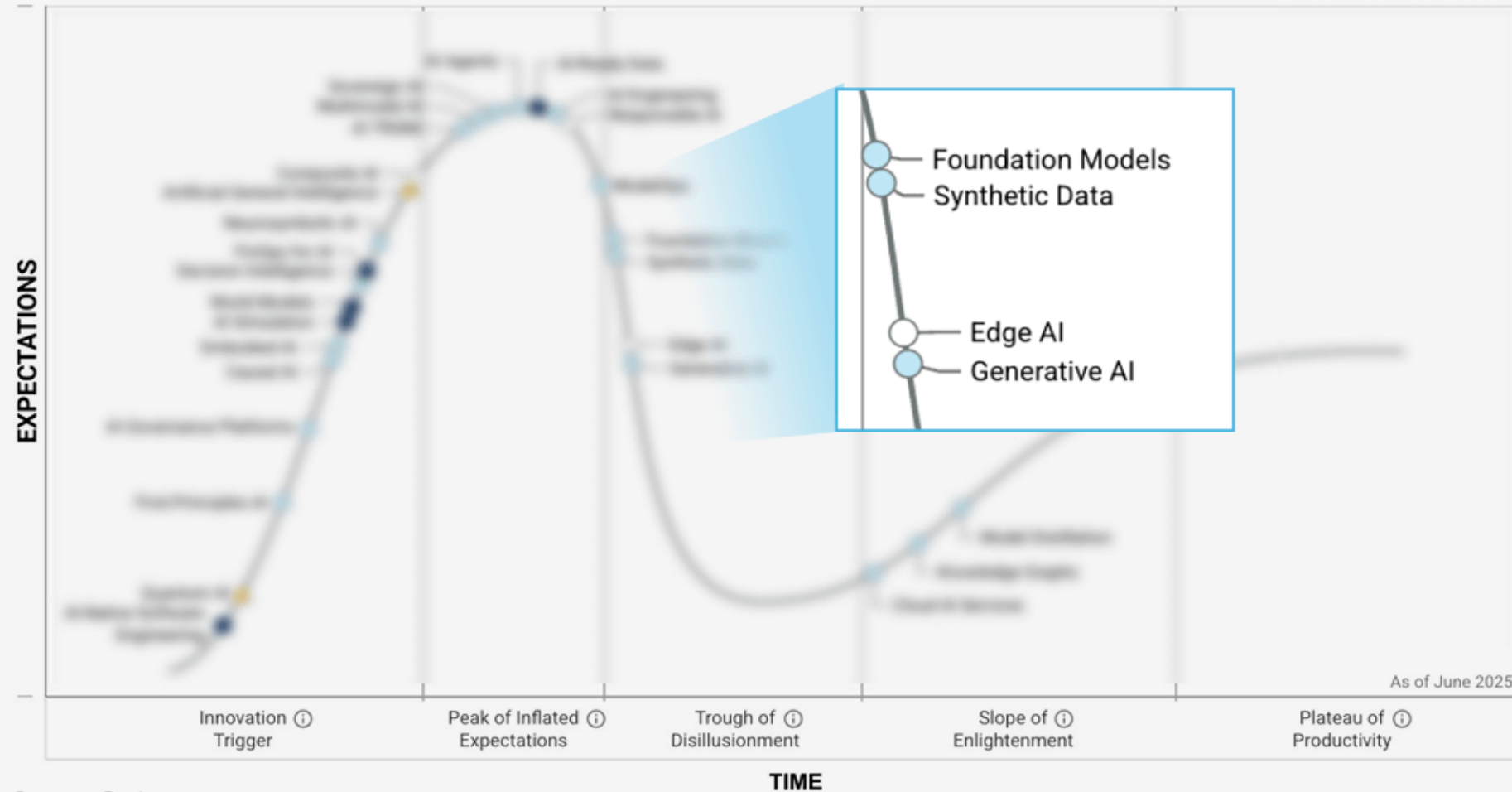
Plateau will be reached:

○ < 2 yrs.

● 2–5 yrs.

● 5–10 yrs.

▲ > 10 yrs.



Source: Gartner

© 2025 Gartner, Inc. and/or its affiliates. All rights reserved. CTMKT_3823654

Gartner



Articles

About 4.850 results (0,09 sec)

Any time

Since 2025

Since 2024

Since 2021

Custom range...

Sort by relevance

Sort by date

Any type

Review articles

☐ include patents

☒ include citations

☒ Create alert

Extremal Testing for Network Software using LLMs

[R Singha](#), [H Qian](#), [S Saikrishnan](#), [T Zhao](#)... - arXiv preprint arXiv ..., 2025 - arxiv.org

... This is because this research is in a context where: a) the **LLM** does not understand the intent of the particular software being tested (unlike in our case for say **DNS** or HTTP); and b) the ...

☆ Save Cite Related articles All 2 versions

[PDF] arxiv.org

Fine-tuning Large Language Models for DGA and DNS Exfiltration Detection

[MA Sayed](#), [A Rahman](#), [C Kiekintveld](#)... - 2024 Annual ..., 2024 - ieeexplore.ieee.org

... LLMs for detecting DGAs and **DNS** exfiltration attacks. We developed **LLM** models and conducted ... Our **LLM** model significantly outperformed traditional natural language processing

☆

[PDF] ieee.org

[PDF]

RC

... in

instr

☆

LLM + DNS

[PDF] deepness-lab.org

[HTML] Rule-Based eXplainable Autoencoder for DNS Tunneling Detection

[G De Bernardi](#), [GB Gaggero](#), [F Patrone](#), [S Zappatore](#)... - Computers, 2025 - mdpi.com

... Machine (**LLM**). The main contribution of this paper is a method for detecting **DNS** tunneling ... of rules by using DT and **LLM**, we consider three different scenarios as outlined in Figure 4. ...

☆ Save Cite Related articles

[HTML] mdpi.com

Volltext UB der UniBW M

AgentDNS: A Root Domain Naming System for LLM Agents

[E Cui](#), [Y Cheng](#), [R She](#), [D Liu](#), [Z Liang](#), [M Guo](#)... - arXiv preprint arXiv ..., 2025 - arxiv.org

... While **DNS** effectively decouples human-readable names from machine-level addressing, ... **LLM** agents require autonomous service discovery and interoperability. Traditional **DNS** lacks ...

☆ Save Cite Related articles All 2 versions

[PDF] arxiv.org

Poster: DoHunter: A feature fusion-based LLM for DoH tunnel detection

[J Diao](#), [S Zhao](#), [J Xie](#), [R Xie](#), [G Shi](#) - Proceedings of the 2024 on ACM ..., 2024 - dl.acm.org

... **DNS** over HTTPS (DoH) reduces the risk of privacy leakage of **DNS** queries, but it also ... context comprehension of Large Language Model (**LLM**) and incorporates expert features to ...

☆ Save Cite Cited by 1 Related articles

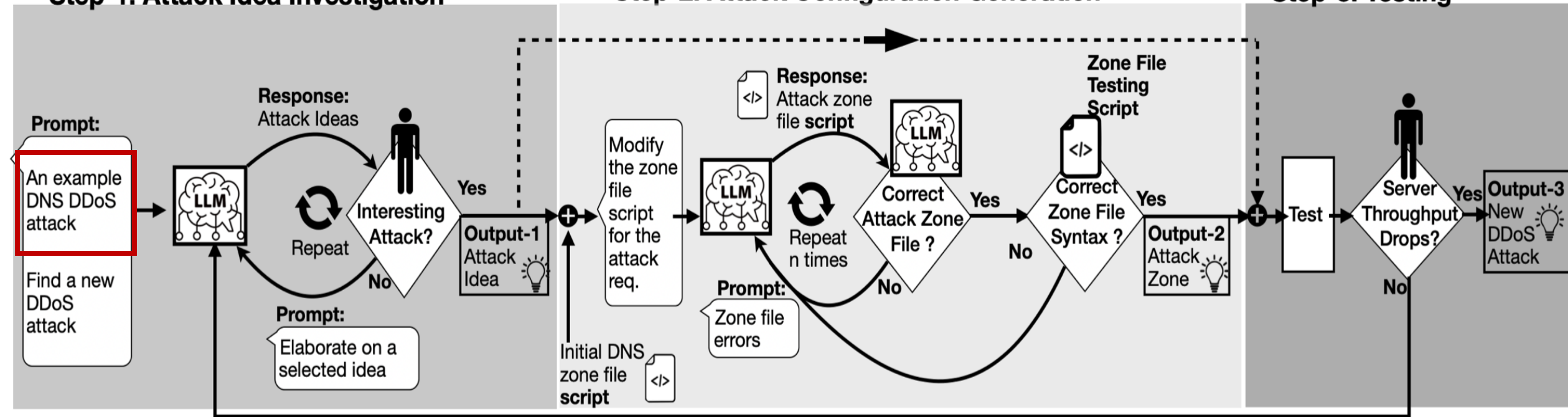
[PDF] acm.org

LLM-Assisted PRotocol Attack Discovery

Step-1: Attack Idea Investigation

Step-2: Attack Configuration Generation

Step-3: Testing



[1] Aygun, R. Can, Yehuda Afek, Anat Bremner-Barr, and Leonard Kleinrock. "LAPRAD: LLM-Assisted Protocol Attack Discovery," *IFIP Networking IOCRCI Workshop*

LLMs for DGA Detection

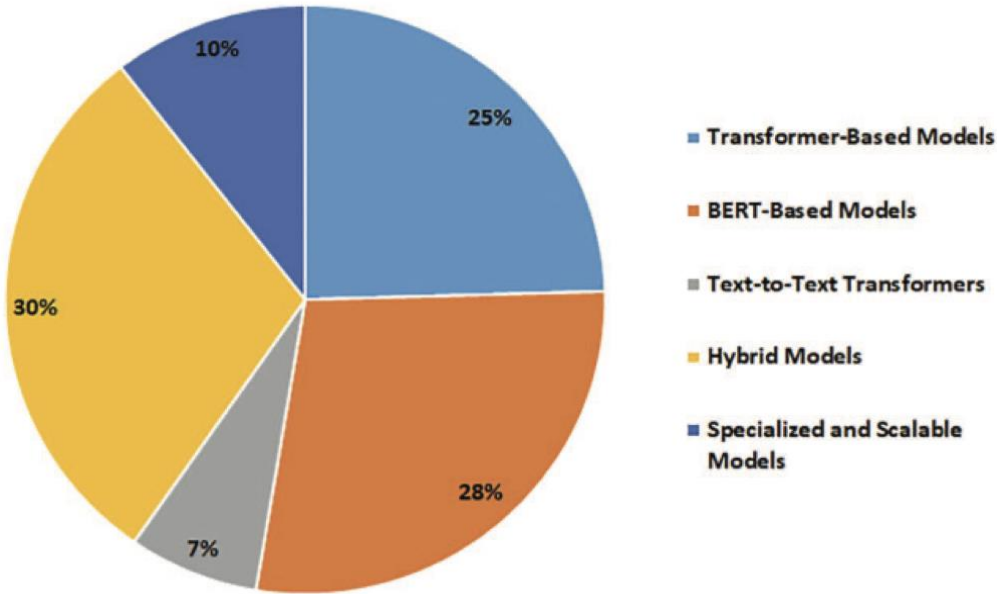


Figure 6: Distribution of LLM model families applied to DGA detection

Table 2: Taxonomy of LLM families for DGA detection with performance and deployment metrics

Model family	Representative models	Key strengths	Limitations	Accuracy	FPR	AUC	Latency
Transformer-based models	GPT-2, GPT-3, LLaMA, RoGPT	Autoregressive learning, strong sequence modeling, zero-shot capabilities	High compute cost, not suitable for real-time edge deployment	94%–96%	3%–5%	0.89–0.92	High
BERT-based models	BERT, RoBERTa, DomURLs-BERT	Bidirectional token understanding, high recall on dictionary DGAs, fine-tunable	Requires task-specific retraining, moderately scalable	92%–97%	2%–4%	0.90–0.94	Medium
Text-to-text transformers	T5, T5-Contrastive Label Generation (CLG), XLNet	Few/zero-shot performance, explainable outputs, flexible text generation	Larger model size, slower inference than encoder-only models	84%–93%	3%–6%	0.85–0.90	Medium-High
Hybrid embedding models	Word2Vec + LSTM, BPBZ, ELMo	Multiscale features, compact for edge, supports limited-resource deployments	Lower semantic depth, precision degradation on novel DGAs	89%–94%	2%–4%	0.87–0.91	Low
Specialized & scalable LLMs	ERNIE, Megatron-LM, Turing-NLG	Scalable, multilingual, high-capacity for cross-domain detection	High training/inference cost, not edge-compatible	92%–95%	3%–5%	0.88–0.91	High

[2] Alqahtani, Hamed, and Gulshan Kumar. “Large Language Models for Effective Detection of Algorithmically Generated Domains: A Comprehensive Review.” *Computer Modeling in Engineering & Sciences* 144, no. 2 (2025): 1439.

About 5.680 results (0.09 sec)

External Testing for Network Software using LLMs

[PDF] arxiv.org

B. Athang H. Gujlabrichan -Tjahok - st-ajanshansen -Tulkeū -, 2020 + ppriv.oww.crg

The external gilrlens phet/hms./nevenified an action/wellher has LLMS *intermtecs the accsption, the* lower approach character optimized nto a dighter way for us use to pay as DNS HTTPi and by....

★ Savro Inc. sqrtabo - Rstalleo artvice disponibles

Fine-tuning Large Language Models for DGA and DNS Extilization Detection

_____n

LLM + DNS

LLM + DNS

LLM + DNS

IMTUAIRule Based eXplainable Autoencoder for DNS Tunneling Detection

[PDF] md] org



Not just: “Let’s try because LLMs are popular.”

Waklebard.org

Preferential Flow Mechanism of the LLM for Doin Tunneling Detection

DNS for LLMs: *Why DNS is a good input?*

- ✓ Symbolic and text-based: naturally fits LLM tokenization

Human-readable, structurally complex

Real-world attack surface with diverse threats

Benchmark for structured reasoning

DNS for LLMs: *Why DNS is a good input?*

Symbolic and text-based: naturally fits LLM tokenization

✓ **Human-readable, structurally complex**

Real-world attack surface with diverse threats

Benchmark for structured reasoning

DNS for LLMs: *Why DNS is a good input?*

Symbolic and text-based: naturally fits LLM tokenization

Human-readable, structurally complex

✓ **Real-world** attack surface with diverse threats

Benchmark for structured reasoning

DNS for LLMs: *Why DNS is a good input?*

Symbolic and text-based: naturally fits LLM tokenization

Human-readable, structurally complex

Real-world attack surface with diverse threats

✓ **Benchmark** for structured reasoning

LLMs for DNS: *Why LLMs are good for DNS?*

✓ Semantic / context detection

Prompt-based Zero / Few-shot generalization

Explainability via natural language

Actionable response generation

LLMs for DNS: *Why LLMs are good for DNS?*

Semantic / context detection

✓ Prompt-based **Zero / Few-shot** generalization

Explainability via natural language

Actionable response generation

LLMs for DNS: *Why LLMs are good for DNS?*

Semantic / context detection

Prompt-based Zero / Few-shot generalization

✓ **Explainability** via natural language

Actionable response generation

LLMs for DNS: *Why LLMs are good for DNS?*

Semantic / context detection

Prompt-based Zero / Few-shot generalization

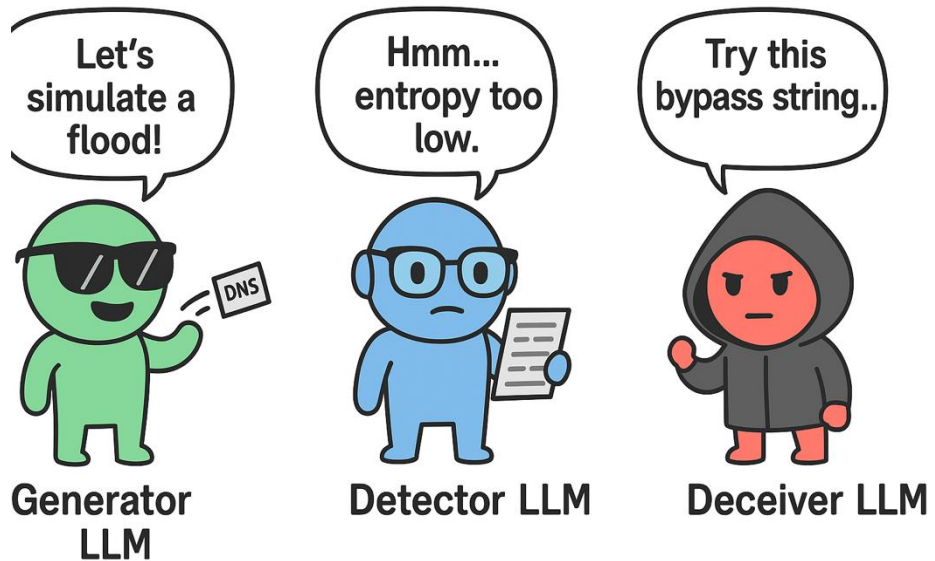
Explainability via natural language

✓ Actionable **response** generation

LLMs vs. Traditional Methods

Aspect	Traditional ML/Statistical Approach	LLM-based Approach
Input	Hand-crafted features, flow statistics	Raw/semi-structured DNS sequences + features
Adaptability	Strong in-domain; Weaker on novel attacks	Prompt-based zero/few-shot; adapts with minimal training
Explainability	Limited (scores, feature weights)	Natural-language explanations, Human-readable reasoning
Attack Coverage	Good for volumetric & known patterns	Broader: flooding, amplification, semantic abuses, policy misuses
Weakness	Efficient but brittle to unseen variants	Higher cost/latency, adversarial risks

LLMs for DNS: *How Can LLMs be effectively used?*

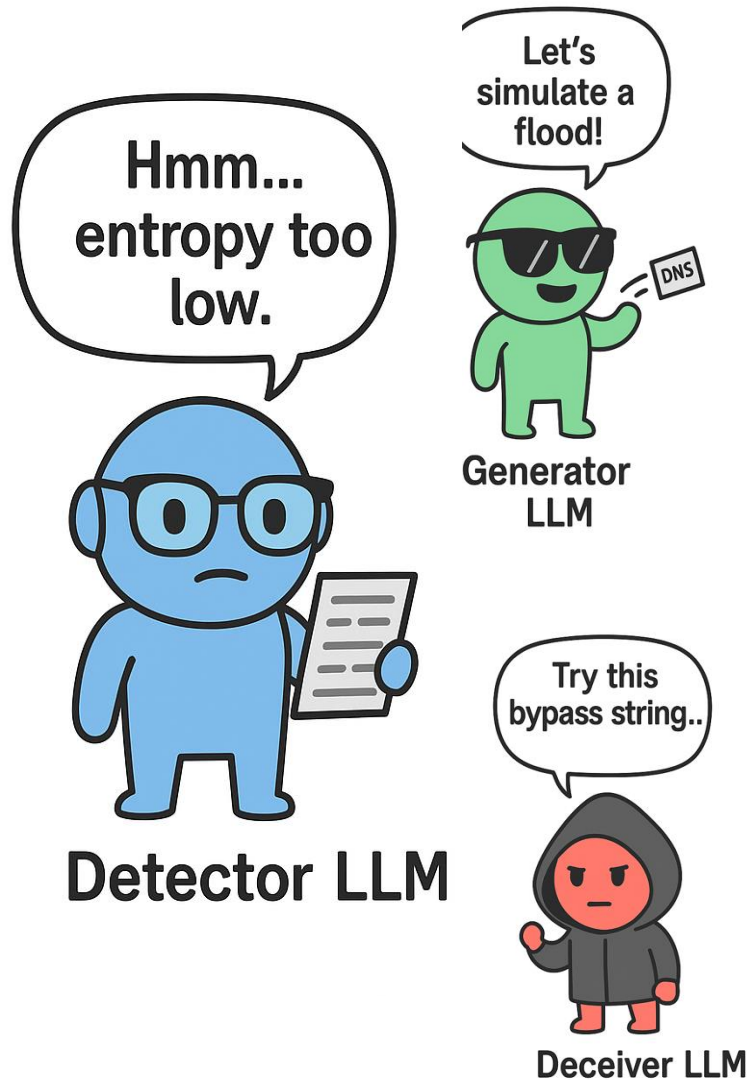


Create Synthetic DNS Attack Traffic

Detect and Explain DNS Attack Traffic

Confuse Detector with Adversarial Traffic

LLMs for DNS: *How Can LLMs be effectively used?*



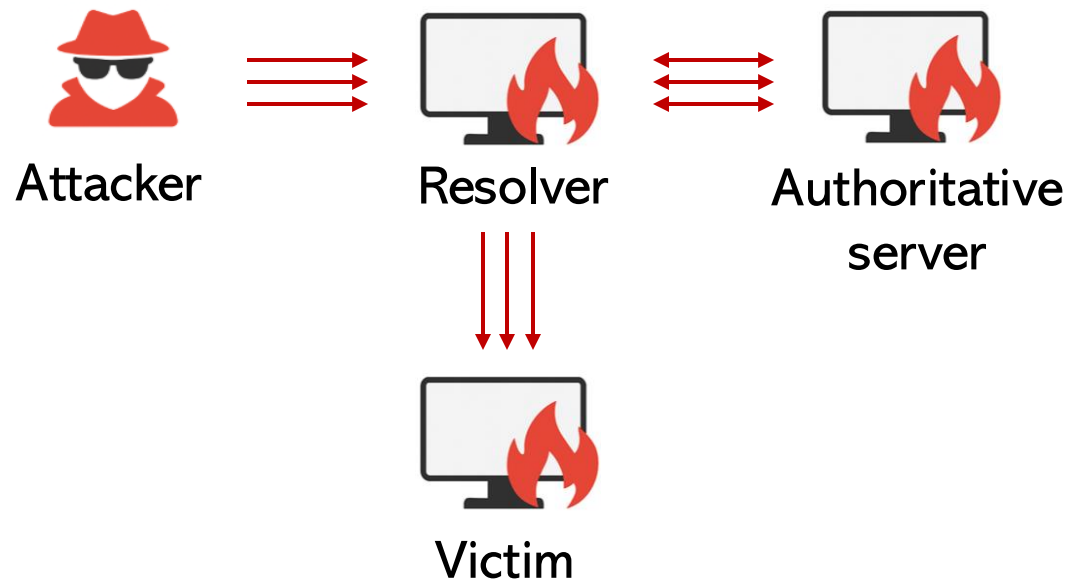
Create Synthetic DNS Attack Traffic

Detect and Explain DNS Attack Traffic

Confuse Detector with Adversarial Traffic

DNS Abuse Taxonomy

DNS as tool or target for DDoS



- ✓ **Flooding**
- ✓ **Reflection / Amplification**
- ✓ **Redirection**
- ✓ **Subversion**
- ✓ **DNSSEC Abuse**

DNS Abuse Taxonomy

Abuse Type	Description	Related DDoS Vector
<u>Flooding</u>	Sending excessive DNS queries to exhaust server resources	Direct DDoS, Resource exhaustion
<u>Reflection</u> <u>/Amplification</u>	Exploiting resolvers to reflect and amplify traffic toward a victim	Amplification DDoS
<u>Redirection</u>	Manipulating DNS responses to redirect traffic (e.g., to a botnet controller)	Indirect DDoS, Traffic Manipulation
<u>Subversion</u>	Compromising domain registration or zone control to manipulate traffic	Indirect DDoS, Traffic Manipulation
<u>DNSSEC Abuse</u>	Abusing DNSSEC's large responses or misconfigurations to overwhelm systems	Amplification DDoS, Resource exhaustion

DNS Abuse Taxonomy

Class	Subclass
<u>Flooding</u>	Query Flooding
	Response Flooding
	NXDOMAIN Flooding (slow drip, random subdomain)
	Resolution Failure Flooding (domain lock-up, phantom subdomain)
<u>Reflection</u> <u>/Amplification</u>	iDNS
	TsuNAME
	Unchained
	NXNS Attack
	NRDelegation Attack
	Loop Attack

Class	Subclass
<u>Redirection</u>	Kaminsky Attack
	DNS Cache Poisoning
	SAD DNS
	Domain Hijacking
	Packet Interception
<u>DNS</u> <u>Subversion</u>	DNS Tunneling
	Fast Flux
	DGA (Malware C2 Infra)
<u>DNSSEC</u> <u>Abuse</u>	DNSSEC Amplification
	NSEC/NSEC3 Walking
	Bogus DNSSEC Data Injection
	Algorithm Downgrade Attack

DNS Abuse Taxonomy

Class	Subclass
<u>Flooding</u>	Query Flooding
	Response Flooding
	NXDOMAIN Flooding (slow drip, random subdomain)
	Resolution Failure Flooding (domain lock-up, phantom subdomain)
<u>Reflection</u> <u>/Amplification</u>	iDNS
	TsuNAME
	Unchained
	NXNS Attack
	NRDelegation Attack
	Loop Attack

Class	Subclass
<u>Redirection</u>	Kaminsky Attack
	DNS Cache Poisoning
	SAD DNS
	Domain Hijacking
	Packet Interception
<u>DNS</u> <u>Subversion</u>	DNS Tunneling
	Fast Flux
	DGA (Malware C2 Infra)
<u>DNSSEC</u> <u>Abuse</u>	DNSSEC Amplification
	NSEC/NSEC3 Walking
	Bogus DNSSEC Data Injection
	Algorithm Downgrade Attack

DNS Abuse Taxonomy

TABLE I
DNS ATTACKS: TECHNICAL AND DETECTION-ORIENTED VIEW.

Category	Attack Name	Key Characteristics	Traffic Features	Traffic Generation	LLM Detectability	Deception Risk
Flooding	Query Flooding	High volume of queries	query_rate↑	Easy	High	Low
	Response Flooding	High volume of responses	response_rate↑	Easy	High	Low
	NXDOMAIN Flooding	Random/non-existent subdomains	RCODE(3)_ratio↑, query_rate_and_entropy↑, unique_qname_count↑, qname_dist_entropy_norm↑	Easy	High	Moderate
	Resolution Flooding	Queries that trigger SERVFAIL/-FORMERR/REFUSED/NOTIMP	RCODE(1,2,4,5)↑, retry_count↑	Easy	High	Moderate
	Slow Drip	Slow transmission of queries/responses to hold resolver threads	retry_count↑	Hard	Low	High
Redirection	Domain Hijacking	Malformed/misleading responses disrupt flow	retry_count↑, RCODE(1,2,4,5)↑, tc_loop↑, empty_noerror_ratio↑	Hard	Low	High
	Kaminsky Attack					
	TXID/port guessing for spoofed responses					
	txid_variety↑, src_port_variety↑, query_burst↑					
Reflection/Amplification						
DNSSEC Abuse						
DNSSEC Amplification						
Signed-record queries → huge responses						
Zone enumeration via proofs of non-existence						
Invalid/bogus DNSSEC records → validation failure						
Preference for weaker validation algorithms						
Validation path weakness						
Very easy						
Moderate-High						
Medium						
Moderate						
Hard						
Hard						
Low						
High						
High						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
Moderate						
		</				

DNS Abuse Detection WorkFlow

From
data
collection
...

RAW (.pcap)

dns_extractor.py

compute_features.py

llm_formatter.py

train_t5.py
predict_t5.py



dns_attack_taxonomy.json

- Class / SubClass
- Explanation
- Feature_Conditions
- Example_Instance

dns_llm_train.json
dns_llm_val.json
> input_text: prompt
> target_text:
structured JSON

app_dns_t5.py

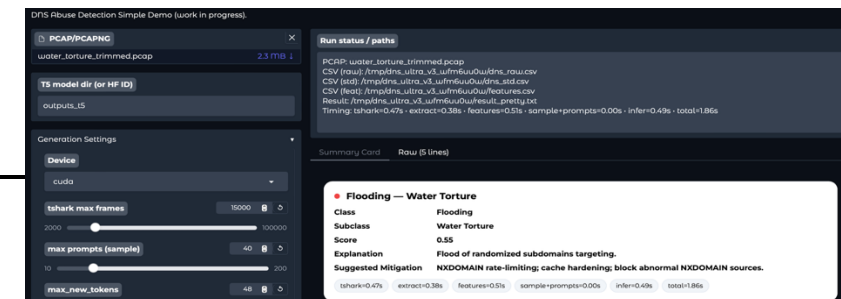
Knowledge Base
(Protocol + Attack knowledge)

DNS RFCs

- [RFC-DNS.pdf \(ODG\)](#) - DNS RFCs (2020-08-29)
- [RFC-DNS-BASIC.pdf \(ODG\)](#) - DNS RFCs - Basic DNS Specification (2020-08-29)

```
test.ipynb X dns_attack X
1 {
2   "Flooding": [
3     {
4       "SubClass": "Query Flooding",
5       "Explanation": "Can be detected by monitoring unusually high query rates and low source
entropy.",
6       "feature_conditions": [
7         {"feature": "query_rate", "condition": "very high (>1000 qps)"},
8         {"feature": "src_entropy", "condition": "low (same source IP or narrow distribution)"}
9       ],
10      "example_instance": {
11        "query_rate": 1800,
12        "src_entropy": 0.3,
13        "query_name": "www.example.com",
14        "query_type": "A",
15        "timestamp": 1735734412.123,
16      },
17    },
18  ]
19 }
```

Gradio Web UI

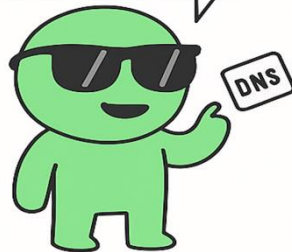


DNS Abuse Detection WorkFlow

Input

Input

- DNS logs/pcaps
- QNAME/QTYPE
- RCODEs
- query rate & intervals
- QNAME entropy
- NXDOMAIN/SERVFAIL
- GeoIP/ASN

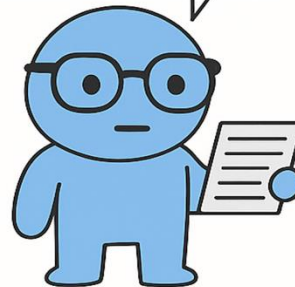


Which datasets/features?

Model

Model

- Feature → prompt
- RAG over playbooks
- few-shot + templates
- rule+LLM fusion
- anomaly scoring
- calibration



Which LLM family/version?

How you'll feed data?

Output

Output

- Alert type & score
- rationale/explanation
- IOCs (ldomains/IPs/ASNs)
- suggested mitigations/queries
- confidence



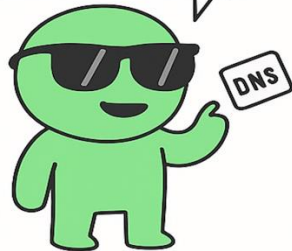
What you want back?

Input Models

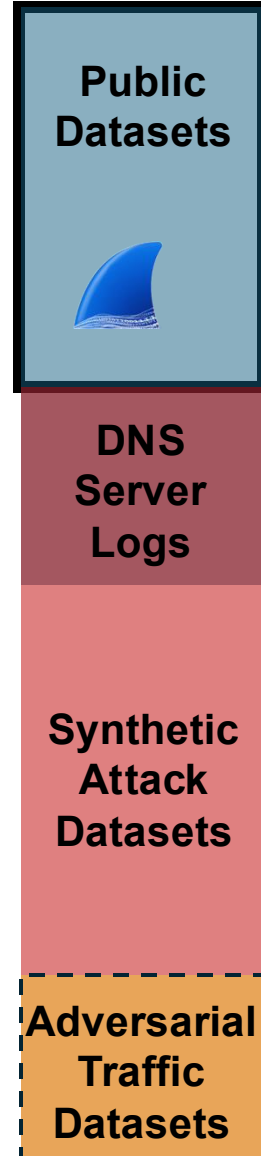
Input

Input

- DNS logs/pcaps
- QNAME/QTYPE
- RCODEs
- query rate & intervals
- QNAME entropy
- NXDOMAIN/SERVFAIL
- GeoIP/ASN



Which datasets/features?



- Benign traffic:
> OPENINTEL datasets (ground baseline)

- Attack traffic:
> Open-source attack datasets

- Traffic Generators + LLM-augmentation
> Flamethrower
> Scapy, TRex, MoonGen

(Optional)

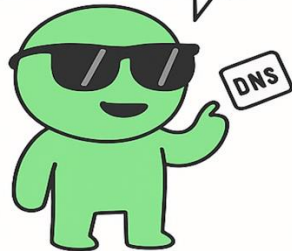
Q. Traffic logs from DNS Operators?

Input Models

Input

Input

- DNS logs/pcaps
- QNAME/QTYPE
- RCODEs
- query rate & intervals
- QNAME entropy
- NXDOMAIN/SERVFAIL
- GeoIP/ASN



Which datasets/features?

Public
Datasets



DNS
Server
Logs

Synthetic
Attack
Datasets

Adversarial
Traffic
Datasets

Synthetic DNS Attack Traffic Generation for Cyber Threat Intelligence with LLM Augmentation

Abstract—DNS abuse is a central component of CTI, as malicious domains and query behaviors are among the most widely shared indicators of compromise. However, real DNS attack datasets remain scarce due to privacy and operational constraints, which hampers reproducible research and systematic evaluation of detection methods. We introduce a framework for *synthetic DNS attack traffic generation*, targeting key abuse scenarios such as flooding, amplification, and redirection. The

capture pcaps and structured logs (dnstap) and compute validation metrics and detection features in reproducible scripts.
- Tool-based backbone (flamethrower, dnssim, Unbound)
- Attack scenarios: Flooding, Amplification, Redirection
- Parameters: rate, QNAME entropy, QTYPE, TTL, delegation depth
- LLM augmentation: domain patterns, evasive variants


- Traffic Generators + LLM-augmentation
 - > Flamethrower
 - > Scapy, TRex, MoonGen

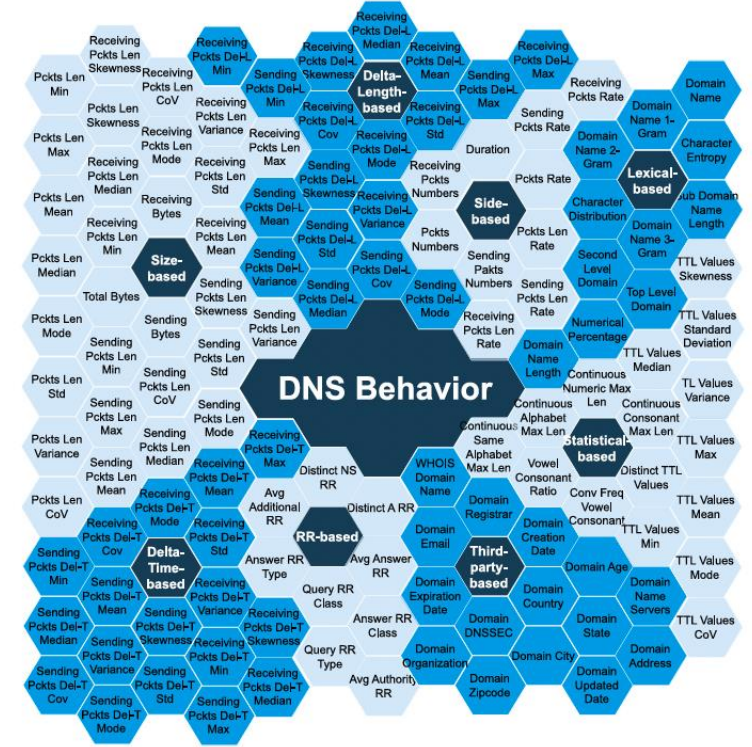
(Optional)

Q. Traffic logs from DNS Operators?

Input

Input

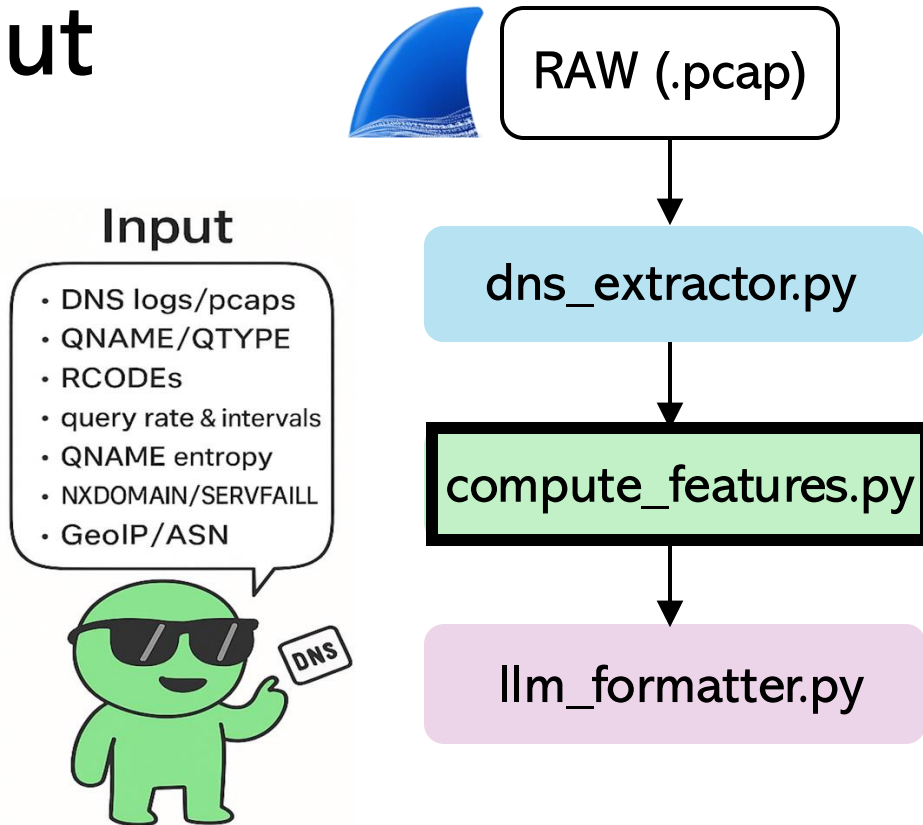
- 
- A green cartoon character with a round head, wearing black sunglasses and a simple green body. The character is holding a white rectangular sign with the text 'DNS' in black capital letters. A small black lightning bolt is above the character's head.



```
frame.time_epoch, ip.src, ipv6.src,  
dns.flags.response, dns.qry.name, dns.resp.name,  
dns.resp.ttl, dns.resp.len, dns.flags.rcode,  
dns.count.answers, dns.count.add_rr,  
udp.length, frame.len
```

Input Models

Input



Which datasets/features?

Outputs for DNS analysis
and LLM prompting:

--mode packet

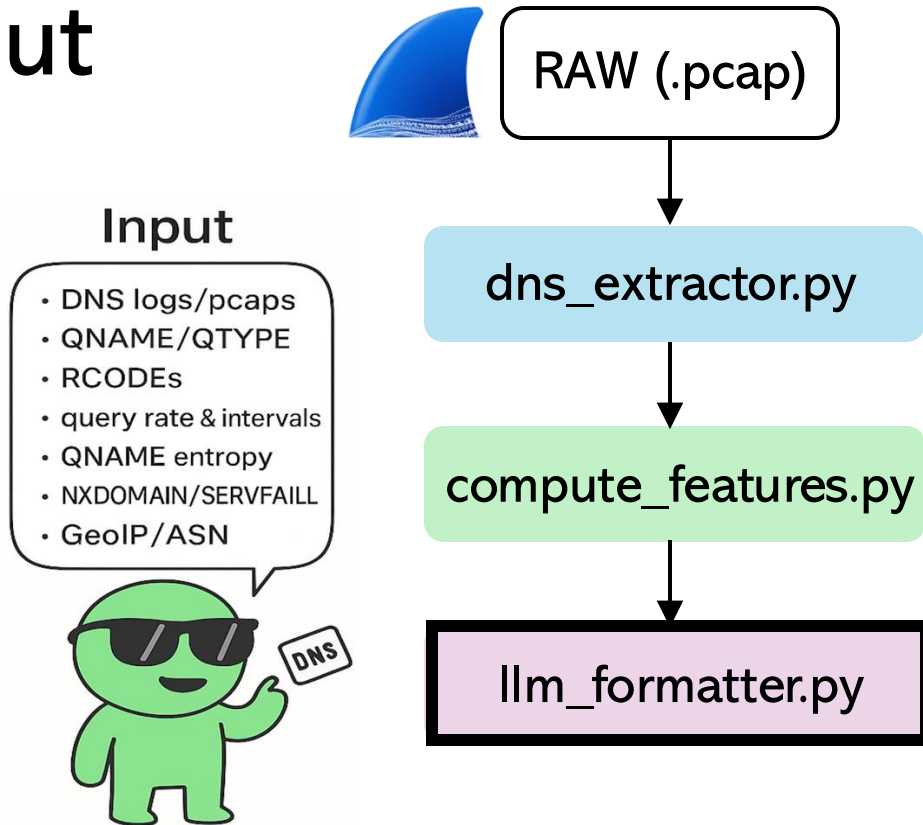
Fine-grained per-packet context
+ file-level aggregates

--mode window

Sliding-window aggregation over
--window seconds;
> Per-window features

Input Models

Input



Which datasets/features?

Converts 'features.csv' into JSONL for T5 training or inference prompts

--mode infer

Build/keep 'input_text,' append schema suffix, and write all prompts to 'out_val'

--mode train

Requires 'label'; performs safe train/val split
Use taxonomy to build 'target_text'

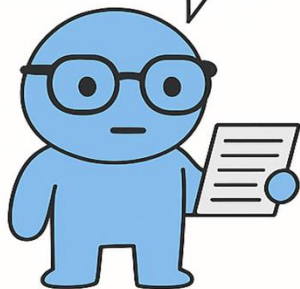
LLM Models

We need “ Classification + Explanation + Generation + ... ”

Model

Model

- Feature → prompt
- RAG over playbooks
- few-shot + templates
- rule+LLM fusion
- anomaly scoring
- calibration



Which LLM family/version?

How you'll feed data?

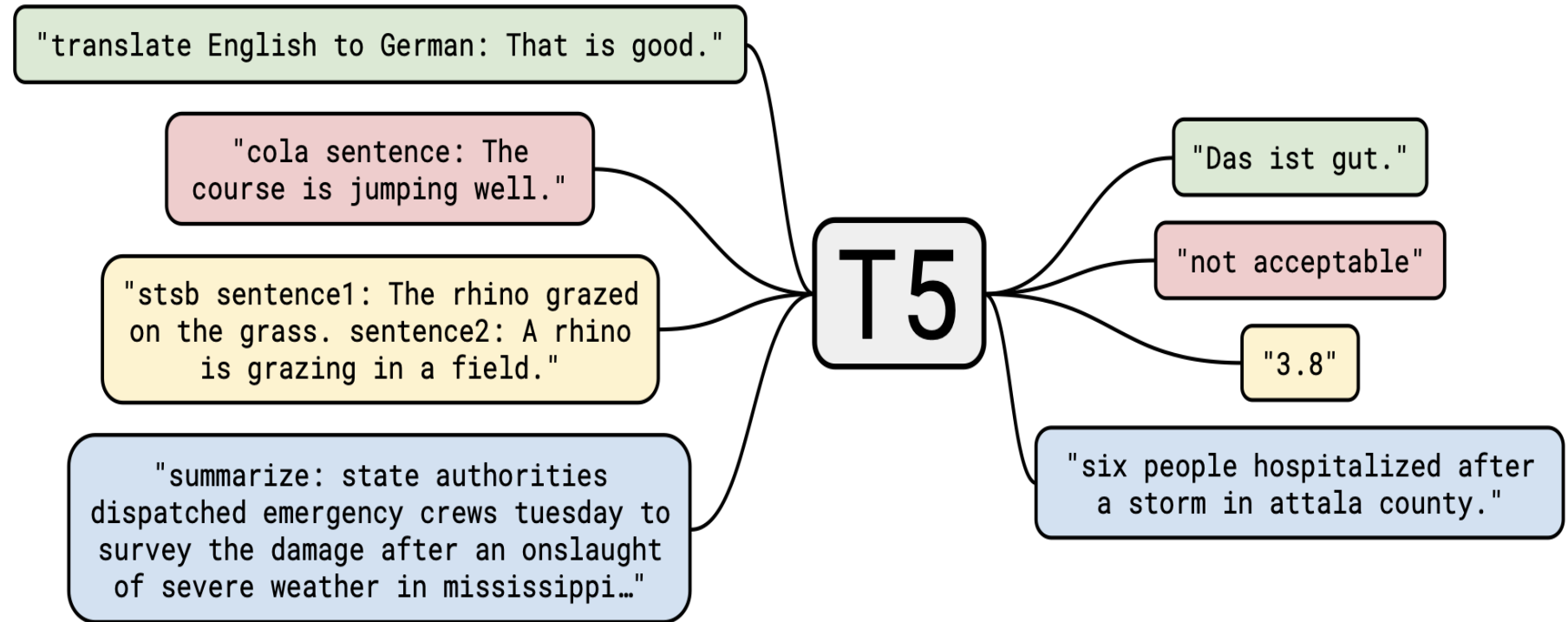
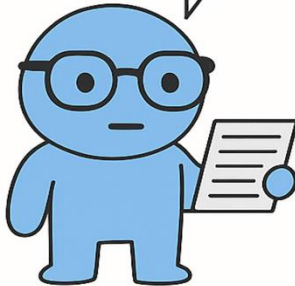
Type	Models	Key Characteristics
Decoder-only	GPT-Family, LLaMA	Strong at log/query generation Autoregressive: predicts next token sequentially Less suitable for classification/detection tasks
Encoder-only	BERT-Family	High accuracy in detection/classification Cannot generate outputs; Limited explanatory power
Encoder-Decoder	T5 , BART, UL2	Supports detection, explanation, and generation Larger parameter size / higher computational cost

LLM Models

Model

Model

- Feature → prompt
- RAG over playbooks
- few-shot + templates
- rule + LLM fusion
- anomaly scoring
- calibration



T5: Text-to-Text Transfer Transformer

Which LLM family/version?

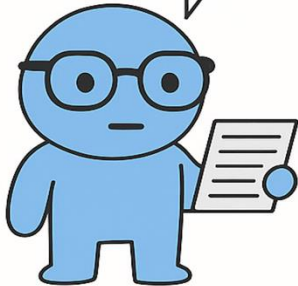
How you'll feed data?

LLM Models

Model

Model

- Feature → prompt
- RAG over playbooks
- few-shot + templates
- rule+LLM fusion
- anomaly scoring
- calibration



FLAN-T5
Model

train_t5.py
predict_t5.py

app_dns_t5.py

T5 Training & Inference Scripts

train_t5.py

Fine-tunes T5 on DNS prompts/targets
with robust metrics

Tokenizer/Model: **AutoTokenizer**,
AutoModelForSeq2SeqLM

predict_t5.py

Batch-generated normalized single-line
JSON per input

Which LLM family/version?

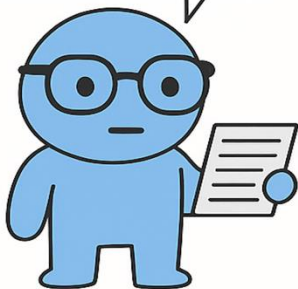
How you'll feed data?

LLM Models

Model

Model

- Feature → prompt
- RAG over playbooks
- few-shot + templates
- rule+LLM fusion
- anomaly scoring
- calibration



Which LLM family/version?

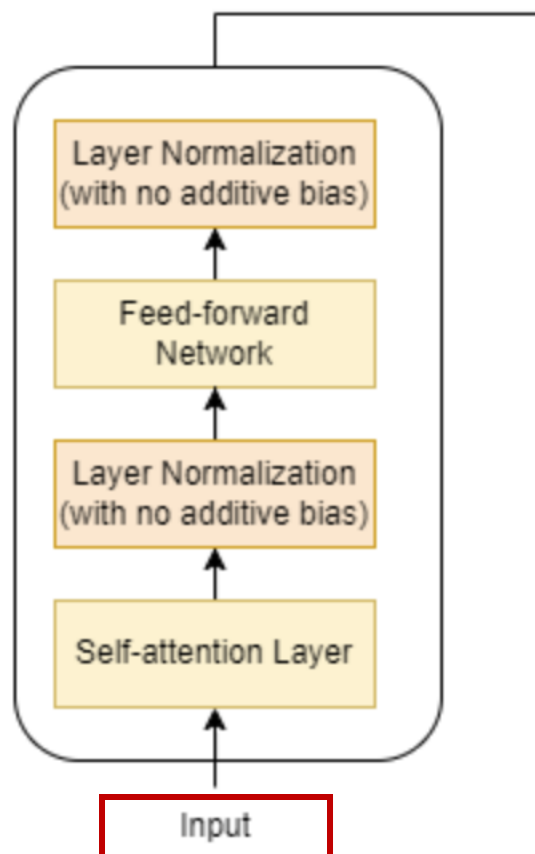
How you'll feed data?

T5-Base Model

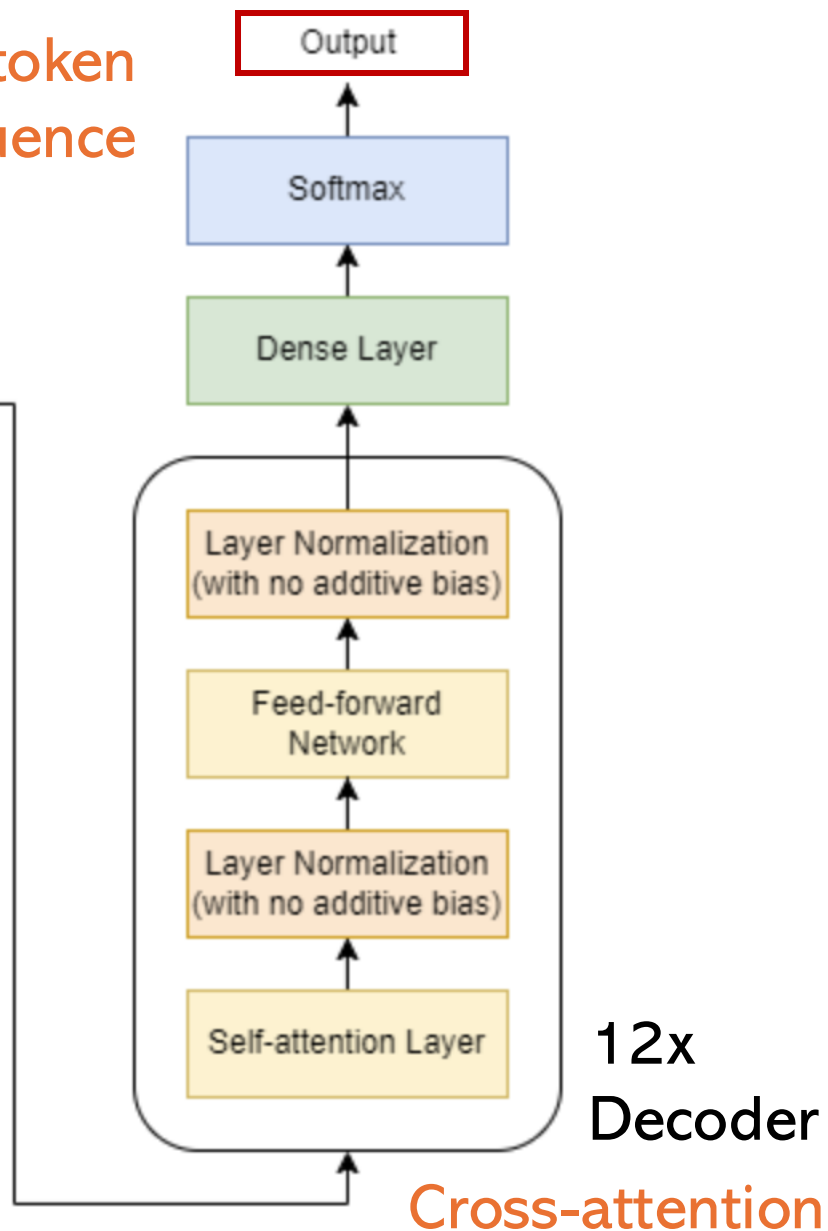
large
small
...

JSON token
sequence

12x
Encoder
Self-attention



Formatted DNS Log

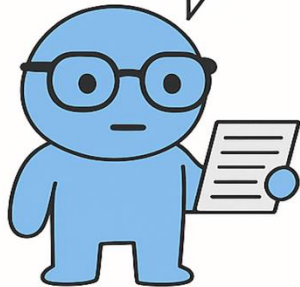


LLM Models

Model

Model

- Feature → prompt
- RAG over playbooks
- few-shot + templates
- rule+LLM fusion
- anomaly scoring
- calibration



Which LLM family/version?

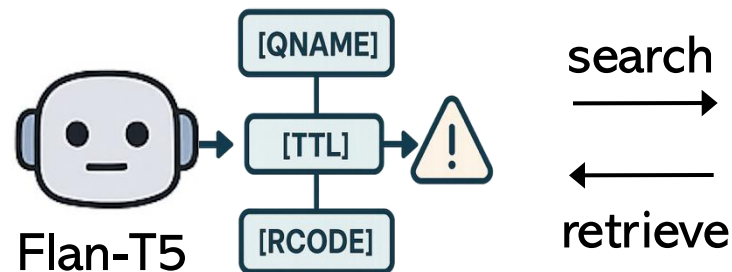
How you'll feed data?

- DNS attack taxonomy for **LLM Inference**:

```
1 {  
2   "Flooding": {  
3     {  
4       "SubClass": "Query Flooding",  
5       "Explanation": "Can be detected by monitoring unusually high query rates and low source  
6       entropy",  
7       "feature_conditions": [  
8         { "feature": "query_rate", "condition": "very high (>1000 qps)" },  
9         { "feature": "src_entropy", "condition": "low (same source IP or narrow distribution)" }  
10      ],  
11      "example_instance": {  
12        "query_rate": 1800,  
13        "src_entropy": 0.3,  
14        "query_name": "www.example.com",  
15        "query_type": "A",  
16        "timestamp": 1735734412.123,  
17      },  
18    }  
19  },  
20}
```

dns_attack_
taxonomy.json

- Knowledge base as **RAG**:
 - > Up-to-date data reasoning from RFCs and attack reports

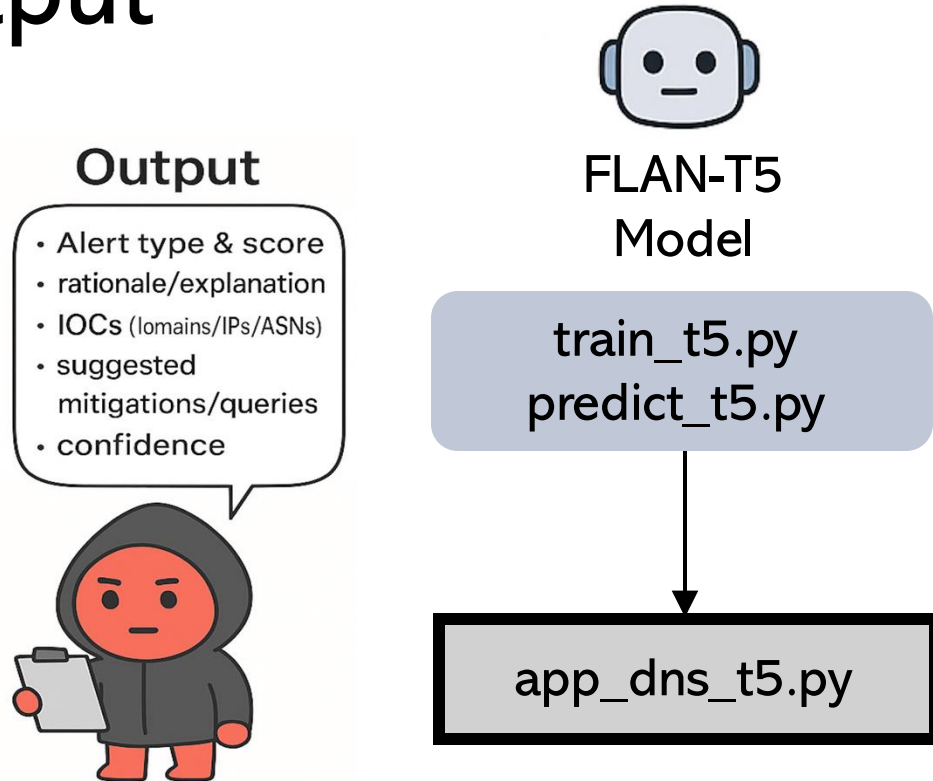


Knowledge Base

kb_rfcs.json
kb_attacks.json

Output Models

Output



What you want back?

PCAP to Inference: End-to-End Demo

- Gradio Web UI

Summary Card (Gradio HTML):

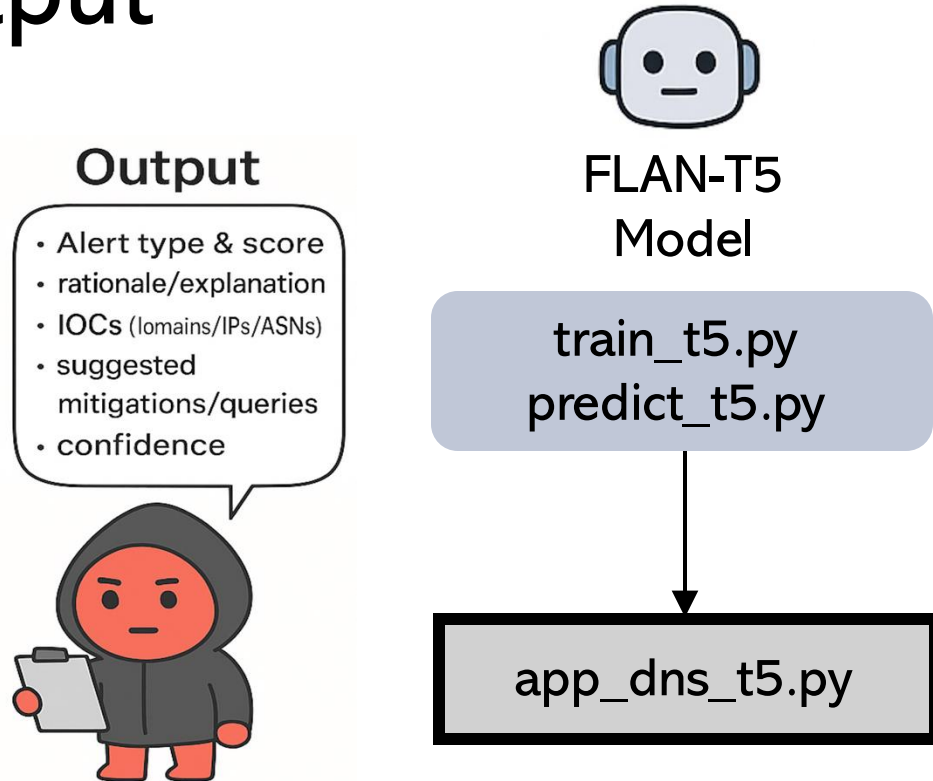
- Class
- Subclass
- Score
- Explanation
- Mitigation

CSV (labelled):

per-row + final aggregated label

Output Models

Output



What you want back?

PCAP to Inference: End-to-End Demo

- Grafana/Prometheus

Prometheus:

- exports/metrics (port: 9108);

Grafana Panels:

- dns_pipeline_seconds
- dns_label_ratio
- dns_final_score
- dns_nxdomain_ratio ...

Future Work

- ✓ **Short-term:** Datasets, Benchmarking
 - > Benchmark diverse DNS datasets
 - > Compare across model families (Traditional MLs vs. LLMs)
- ✓ **Mid-term:** Develop a Synthetic / Adversarial Traffic Framework
 - > Adversarial Robustness Testing
 - > Improve Trustworthiness
- ✓ **Long-term:** Towards Operator-grade Deployment

Simple DEMO (3m)