

Summary: Applying NAIRR Pilot Resources to Optimize AI-Ready Data Investments

Overview

We propose an EAGER project to demonstrate how NAIRR Pilot resources at SDSC can facilitate the development of a new service to the cybersecurity research community, specifically to **infer the utility of datasets and software tools** based on their documented use in scientific publications. This solution will require data collection, preprocessing, model selection/training, and evaluation/deployment. The service will address persistent community challenges, and serve as a model for related problems in engineering as well as social science disciplines.

The service capability will **assess scientific data utility to optimize AI-ready data investments**. For many years CAIDA has tracked use of its scientific data in publications, in part to support its rich context catalog that links research papers to resources (data sets and software tools) used in these papers. **However, even current state-of-the-art techniques for automatically extracting links between resources cannot handle the complexity, variability, and contextual nuances of natural language**, e.g., in discerning a reference to a resource from its actual use in a paper.

Intellectual Merit

This task will require advanced, context-aware extraction of metadata from unstructured text and associated inference of relevant annotations. We will leverage LLMs to enable more accurate interpretation and explanation of complex relationships in scientific publications. We will combine CAIDA's (cybersecurity and critical network infrastructure) domain-specific expertise with SDSC's AI and cyberinfrastructure expertise to develop LLM-based approaches such as prompt engineering and fine tuning for extracting relevant information from text.

Broader Impacts

Macroscopic Internet data sets are notoriously expensive to collect, curate, maintain, and share; NAIRR can play a key role in navigating these challenges, and lowering the barrier for under-resourced researchers to discover powerful and relevant cybersecurity data and associated research. The proposed work addresses two commonly articulated needs: getting started with Internet data, and understanding the utility of such data for further research. More broadly, the tools and approach will be useful for other data-focused disciplines, addressing what will be a growing need in the AI research community: evaluating the return-on-investment in collecting and curating complex AI-enabled as well as AI-enabling data sets.

EAGER: NAIRR Pilot: Applying NAIRR Pilot Resources to Optimize AI-Ready Data Investments (DCL 24-093)

1 Objective: Assessing Scientific Data Utility

We propose a NAIRR Pilot EAGER project to demonstrate how NAIRR Pilot resources at SDSC can facilitate a new service to the cybersecurity research community: **inferring utility of datasets and software tools** based on their documented use in scientific publications. This solution will require context-aware extraction and interpretation of complex relationships from unstructured text. The service will address persistent community challenges, and will serve as a model for other scientific disciplines that face similar data integrity and data utility assessment challenges.

1.1 Motivation

Sustaining data collection and curation is expensive, and AI-ready data sets will be even more expensive to sustain and curate. Funding agencies need mechanisms to assess which data sets significantly contribute to scientific discovery and innovation. For many years CAIDA has tracked the use of our scientific data in publications, in part to support our rich context catalog that links networking and Internet security research papers to data sets and software tools used in these papers. This catalog has accelerated scientific discovery and development of cybersecurity data science and research skills [1]. **However, even state-of-the-art techniques for automatically extracting this information from publications cannot handle the complexity, variability, and contextual nuances of scientific language**, e.g., in discerning a passing reference to a resource from its actual use in a paper. The problem is exacerbated by the proliferation of research publications, some of which may now be authored by LLMs unbeknownst to their readers. Addressing these limitations requires advanced, context-aware extraction methods, such as those enabled by LLMs, which can leverage patterns in language to interpret complex relationships in scientific texts.

1.2 Why the Proposed Work is Appropriate for EAGER Consideration

Our proposal aims to develop a novel service that infers the utility of datasets and software tools based on their documented use in scientific publications. This initiative aligns with the NSF's EAGER program, which supports exploratory work in its early stages on untested but potentially transformative research ideas or approaches. The proposed research involves untested methodologies for context-aware extraction and interpretation of complex relationships from unstructured text. By leveraging advanced techniques, we aim to overcome the limitations of current methods in discerning nuanced references within scientific literature. This approach embodies the **high-risk, high-reward paradigm** that the EAGER program seeks to promote.

The project combines expertise from cybersecurity, data science, and AI to address persistent challenges in assessing data utility. By leveraging LLMs to extract and interpret complex relationships from unstructured text, the project applies new methodologies to the field of cybersecurity research. This **interdisciplinary approach** is characteristic of the type of research EAGER aims to support. If successful, this project could significantly enhance the way funding agencies and researchers assess the contribution of datasets to scientific discovery and innovation. By providing a model applicable to related issues in other data-focused scientific disciplines, the project has the **potential to transform** current practices in data utility assessment, leading to more informed investment decisions and optimized AI-ready data curation.

1.3 Technical Approach

We propose to leverage CAIDA’s (cybersecurity and network infrastructure) domain-specific expertise and extensive domain-specific datasets, SDSC’s AI and cyberinfrastructure expertise, open-source LLMs, and the NAIRR Pilot resources at SDSC (Expanse and Voyager) to design and prototype a service to **detect, validate, and characterize use of data sets and software tools in scientific publications**.

We will use state-of-the-art open-source LLMs for this task. Specifically, we plan to use LLaMA 3.1, which has a context length of 128K tokens, and has been shown to be competitive with leading open and closed LLMs for a range of tasks [2]. A LLaMa 3.1-70B model [3] is available as part of SDSC LLM, an LLM-as-a-service, currently available to SDSC staff for development purposes. Members of our team initiated the effort to build SDSC LLM as an internal resource to provide quick, cost-effective, and private access to LLM capabilities for SDSC researchers. SDSC LLM uses vLLM [4] for LLM serving and OpenWebUI [5] for chat UI and API access to the underlying models. The LLaMa 3.1-70B model is currently deployed on 4x A100s 80GB GPUs. We will leverage SDSC LLM for tasks based on inference. For tasks requiring model parameter updates, we will make use of AI-optimized accelerators on Voyager and GPUs on Expanse. The Voyager supercomputer, hosted at SDSC, is designed for deep learning workloads [6]. SDSC also hosts the Expanse supercomputer, which provides compute resources for a wide range of applications [7].

We will explore various approaches to leverage LLMs for this task. We will start with prompt engineering, followed with fine tuning, and if time permits, RAG and prompt tuning. Prompt engineering is the process of designing and refining prompts to elicit desired responses from LLMs. We will evaluate several prompt engineering techniques, including prompt crafting, shot prompting, and chain-of-thought. RAG is a method to enhance the quality of LLM responses by incorporating additional information from an external source. Prompt tuning is a technique to adapt LLMs to a specific task by adjusting only a few parameters corresponding to the prompt embeddings to guide the model’s output. We will pursue these latter techniques (RAG and prompt tuning) if time allows, or integrate them into follow-on work.

2 Methodology for inferring resource (data and tools) utility analytics

The objective of this task is to extract knowledge about data and software resources used in Internet research publications. To facilitate human audit of the result, we will have the LLM excerpt sentences from each publication (and citations within these sentences) that indicate use of a specific resource. We will then include text from the cited references to guide and/or fine-tune the LLM. Minimizing false positives will require including negative examples, i.e., publications that reference but do not actually use a specific data set or tool, in the multi-shot prompt. Using manually annotated papers, we will investigate LLM-based approaches for extracting information related to data and software resources from text.

2.1 Preliminary Approach and Example

The three-step prompt engineering approach is typically: (1) create the training data set, (2) use the training data to engineer a set of one-shot prompts using segments of the paper, then a final prompt using the entire paper, and (3) validate the results. We describe the process and then illustrate it with an example analysis for a single publication.

1. Step 1: Create training data set.

We will first manually identify resources (data sets or software tools) used or referenced by a set of research papers. We will represent each resource in YAML with five fields: (Figure 1):

- resource identifier URL if provided in the paper
- resource type (dataset or software)
- classification as merely mentioned or used by the paper
- sentences the human considered necessary to make the labels
- resource’s bibliographic references cited in those sentences.

mentioned	used
<p>While all these works were limited in scope, or only tested specific ECN features, Bauer et al. [22] conducted extensive ECN measurements in 2011 on the Alexa Top 1M list. They found 15 % of TCP-based webservers to be ECN capable, but most hosts in peer-to-peer networks (95 %) to ignore ECN. Using different vantage points, they further identified ~17 % of paths to impair ECN, i.e., change or strip the ECT codepoints and ECN signals.</p> <p>Extending these results, Kühlewind et al. [2] conducted ECN studies in 2012 via research and mobile networks. For the Alexa Top 100k list, they found ~27 % of webservers to support ECN. While >90 % of those correctly mirrored ECN signals when contacted via the research network, no ECN feedback was observed in the mobile network which was attributed to performance proxies and middle-boxes stripping ECN signals. Additionally, the authors conducted the first measurements using IPv6, finding low IPv6 support (less than 2.5 k hosts), but with higher ECN support (48 %). In later</p>	<p>most QUIC stacks on the QUIC interop runner¹ (11 out of 16) currently lack ECN support by design. Joining this observation with the strong ECN support in TCP, it is unclear whether QUIC improves the web w.r.t. ECN support – a gap we close with our study.</p>
<p>4 METHODOLOGY</p> <p>We perform large-scale measurements to assess the usage of ECN in the modern web. First, we identify hosts of the web landscape for which we then assess ECN support. Additionally, we analyze if any irregularities occur in the network. Our overall measurement pipeline is shown in Figure 2. It is fueled by different domain lists, among which we use domain toplist compiled from the (1) Alexa Top 1M¹ [5], (2) Cisco Umbrella [2], (3) Majestic Million [17], and (4) the Tranco Research List [36]. We update these toplist every</p>	
<p>¹Please note that the Alexa Toplist is deprecated from May 2022 on and stopped changing in February 2023</p>	
<p>- used by authors: mentioned sentences: While all these works were limi</p> <p>- used by authors: mentioned sentences: For the Alexa Top 100k list, they</p>	<p>- identifier: Alex Top 1M type: database used by authors: used url: web-archive-org/web/2022010 sentences: It is fueled by different domain li references: - [5] 2022. Keyword Research, Competit</p>

Figure 1: Example of extraction of metadata (and YAML encoding underneath) that indicates a mention of a data set (left) and an actual use of the data set (right). This example is from the publication “ECN with QUIC: Challenges in the Wild” [8].

2. Step 2: Use training data to engineer prompt.

This step will construct a series of examples to use in a multi-shot query [9].

(a) Step 2.1: Creating query-response pairs for a multi-shot example.

We will create a series of query-response pairs each generated from the labeled resource record that we manually extracted from a single paper in Step 1. Each example consists of two pieces: the user prompt (query), and the manually constructed expected response. This pair simulates an exchange between the user and the LLM. As a starting point, the example query will include only sentences that refer to the the labeled resource, and references nearby in the text. Figure 2 provides an example with the user (**role:** user) message containing the example query and the assistant’s (**role:** assistant) message containing the expected response. Note that the ideal one-shot query could include the full text of the labeled paper, but given the LLM’s context length limit, using

the entire paper would limit the number of possible multi-shots.

(b) **Step 2.2: Use examples from Step 2.1 to create a multi-shot query.**

We will include several examples in the input to generate a multi-shot prompt that we provide to the model. The purpose of multi-shot prompting is to provide the model with a set of example queries and expected responses before providing the target query. We will investigate trade-offs between accuracy and processing time with different numbers of shots (i.e., examples).

(c) **Step 2.3: Append final query to multi-and execute final query.** After we construct the set of multi-shot examples, we will append to it a query that includes the full text of the paper. The prompt will end with a single user message with the target paper’s full text. Figure 3 contains an example target query with the full text of the target paper.

3. **Step 3: Sanitize and evaluate the response.**

We will use a tool such as *json_repair* [12] to fix the response format in order to verify if the reference provided in the response is correct. We will identify incorrect responses by finding resources with either a sentence or reference not included in the paper’s full text. Using the set of manually labeled papers, we will evaluate prompt performance by comparing the manually labeled resources against those provided in the LLM’s response.

In addition to prompt engineering, we will also investigate approaches using fine-tuning, and if time permits, RAG and prompt tuning, and compare approaches based on accuracy, processing time, and amount of labeled data required.

2.2 Data to be used

We have already indexed many (a subset of) previous external publications that describe the use of CAIDA datasets. Some of CAIDA’s datasets are already annotated using the annotation schema designed for this (MSRI GMI) project [13], which includes labels such as *AutonomousSystem* [14], and metadata stored in YAML format. We will leverage this previous work to accelerate training.

2.3 Outcome

This task will achieve two critical goals: enabling discovery of the most generative data sets that CAIDA produces, and facilitating a quantitative assessment of the utility of our data sets, which informs investment decisions about which data collection and curation is most important to sustain. Importantly, the tools and approach will be useful for other data-focused disciplines, addressing what will be a growing need in the AI research community: evaluating the return-on-investment in complex AI-enabled as well as AI-enabling data sets.

3 Intellectual Merit

This task will require advanced, context-aware extraction of metadata from unstructured text and associated inference of relevant annotations. We will leverage LLMs to enable more accurate interpretation and explanation of complex relationships in scientific publications. We will combine CAIDA’s (cybersecurity and critical network infrastructure) domain-specific expertise with SDSC’s AI and cyberinfrastructure expertise to develop LLM-based approaches such as prompt engineering and fine tuning for extracting relevant information from text.

```

{
  "role": "user",
  "content": "List datasets or softwares resources that were referenced in the following paper's text. For each dataset or software mentioned in the paper, classify the resource as either merely mentioned or actually used to perform the research described in the paper.

For each resource, print out in JSON format:
- identifier: The resource's identifier.
type: The resource's type, either dataset or software.
used_by_authors: Describes if the resource is used to perform the paper's research or merely mentioned.
url: The resource's URL or (not provided)
sentences: The sentences that infer the resource was used and or merely mentioned in the paper.
references: a list of references used by the sentences
Text:
"""
For the VP selection algorithm, we use the ISI hitlist [25] to select the three representatives of the /24 prefix of each target with the highest score of responsiveness. The authors used the Geonames [2] publicly available service to perform the reverse geocoding queries and obtain the websites that can serve as landmarks if they pass the tests of being locally hosted.
REFERENCES:
[25] Xun Fan and John Heidemann. Selecting representative IP addresses for internet topology studies. In Proc. ACM IMC, 2010.
[2] Geonames API. www.geonames-org/export/ws-overview.html.
""",
},
{
  "role": "assistant",
  "content": '[{
    "identifier": "ISI|IPv4 Hitlist",
    "type": "dataset",
    "used_by_authors": "used",
    "url": "ant-isi-edu/datasets/ip_hitlists/",
    "sentences": "For the VP selection algorithm, we use the ISI hitlist [25] to select the three representatives of the /24 prefix of each target with the highest score of responsiveness.",
    "references": [
      "[25] Xun Fan and John Heidemann. Selecting representative IP addresses for internet topology studies. In Proc. ACM IMC, 2010."
    ],
    "identifier": "Geonames",
    "type": "database",
    "used_by_authors": "merely mentioned",
    "url": "www.geonames-org/export/ws-overview.html.",
    "sentences": "The authors used the Geonames [2] publicly available service to perform the reverse geocoding queries and obtain the websites that can serve as landmarks if they pass the tests of being locally hosted.",
    "references": [
      "[25] Xun Fan and John Heidemann. Selecting representative IP addresses for internet topology studies. In Proc. ACM IMC, 2010."
    ]
  }]'
}

```

Figure 2: JSON representation of example query and expected response that constitute a single shot, generated from “Replication: Towards a Publicly Available Internet scale IP Geolocation Dataset”[10]. The blue text is the target query, green the label resources’ sentences, and red the labeled resources’ references.

```

{
  "role": "user",
  "content": "List datasets or softwares resources that were referenced in the following paper's text. For each dataset or software mentioned in the paper, classify the resource as either merely mentioned or actually used to perform the research described in the paper.

For each resource, print out in JSON format:
- identifier: the resource identifier
type: dataset or software
used_by_authors: true or false, true if the resource was ..
url: the URL for the dataset or software, or (not provided)
sentences: a list of sentences used to infer this information
references: a list of references used by the sentences
Text:
""""
Using Gaming Footage as a Source of Internet Latency Information

Catalina Alvarez
EPFL
Switzerland
catalina.alvarezinostroza@epfl.ch

Katerina Argyraki
EPFL
Switzerland
katerina.argyragi@epfl.ch

ABSTRACT

Keeping track of Internet latency is a classic measurement problem. ....

```

Figure 3: JSON representation of the final message in the prompt, with query in blue and target's full text in purple. (Publication: "Using Gaming Footage as a Source of Internet Latency Information" [11]).

4 Broader Impacts

1. Services to Benefit the Research and Cybersecurity Community:

The ultimate goal of this research is to improve the quality of data and services that CAIDA provides to the research community. Macroscopic Internet data sets are notoriously expensive to collect, maintain, and share; NAIRR can play a key role in navigating these challenges, by using AI tools to inform data utility assessments. The immediate outcome will address two needs commonly articulated by Internet researchers: getting started with Internet data, and understanding the utility of such data for further research.

2. Extension of NAIRR Pilot Capabilities:

This project will extend NAIRR Pilot's capabilities by integrating advanced AI tools for metadata extraction and annotation specific to our domain. We will also provide instructions and code we used, for adaptation by other disciplines.

3. Novel AI-Cyberinfrastructure Innovations:

The development of LLM-based methodologies for metadata extraction from natural language sources represents a significant innovation in AI-cyberinfrastructure. We will update our existing data process pipeline to include these innovations, and document these changes to help other fields benefit from them. Understanding which datasets are most scientifically generative as a whole will help not only scientists, but funding agencies who must set investment priorities.

References

- [1] CAIDA, “Internet Science Resource Catalog,” 2024. <https://catalog.caida.org>.
- [2] Meta, 2024. <https://ai.meta.com/blog/meta-llama-3-1/>.
- [3] Hugging Face, 2024. <https://huggingface.co/meta-llama/Meta-Llama-3.1-70B-Instruct>.
- [4] Woosuk Kwon and Zhuohan Li and Siyuan Zhuang and Ying Sheng and Lianmin Zheng and Cody Yu and Joey Gonzalez and Hao Zhang and Ion Stoica , “vLLM: Easy, Fast, and Cheap LLM Serving with PagedAttention.” <https://docs.vllm.ai/en/latest/>.
- [5] OpenWebUI. <https://openwebui.com>.
- [6] San Diego Supercomputer Center, “Voyager User Guide (NAIRR resource),” 2024. https://sdsc.edu/support/user/_guides/voyager.html.
- [7] San Diego Supercomputer Center, “Expanse User Guide (NAIRR resource),” 2024. https://sdsc.edu/support/user/_guides/expanse.html.
- [8] C. Sander, I. Kunze, L. Blöcher, M. Kosek, and K. Wehrle, “Ecn with quic: Challenges in the wild,” in *Proceedings of the 2023 ACM on Internet Measurement Conference, IMC '23*, p. 540–553, ACM, Oct. 2023.
- [9] “Multi-Shot (multiple examples),” 2023. <https://guide.teahouseai.com/teahouseai/master-llms/main-concepts/multi-shot-multiple-examples>.
- [10] O. Darwich, H. Rimlinger, M. Dreyfus, M. Gouel, and K. Vermeulen, “Replication: Towards a publicly available internet scale ip geolocation dataset,” in *Proceedings of the 2023 ACM on Internet Measurement Conference, IMC '23*, (New York, NY, USA), p. 1–15, Association for Computing Machinery, 2023.
- [11] C. Alvarez and K. Argyraki, “Using gaming footage as a source of internet latency information,” in *Proceedings of the 2023 ACM on Internet Measurement Conference, IMC '23*, (New York, NY, USA), p. 606–626, Association for Computing Machinery, 2023.
- [12] S. Baccianella, “JSON Repair: to repair invalid JSON, used to parse the output of LLMs,” 2024. https://github.com/mangiucugna/json_repair.
- [13] B. Huffaker and k. claffy, “Annotated Schema: Mapping Ontologies onto Dataset Schemas,” tech. rep., Center for Applied Internet Data Analysis (CAIDA), May 2023.
- [14] CAIDA, “CAIDA Ontology.” <https://catalog.caida.org/ontology>.