
Secu-Table: a Comprehensive security table dataset for evaluating semantic table interpretation systems

Azanzi Jiomekong

Department of Computer Science, University of Yaounde I
TIB Leibniz Information Centre for Science and Technology

Jean Bikim

Department of Computer Science, University of Yaounde I
TIB Leibniz Information Centre for Science and Technology

Patricia Negoue

Department of Computer Science, University of Yaounde I

Joyce Chin

Department of Computer Science, University of Yaounde I

Abstract

Evaluating semantic tables interpretation (STI) systems, (particularly, those based on Large Language Models- LLMs) especially in domain-specific contexts such as the security domain, depends heavily on the dataset. However, in the security domain, tabular datasets for state-of-the-art are not publicly available. In this paper, we introduce Secu-Table dataset, composed of more than 1500 tables with more than 15k entities constructed using security data extracted from Common Vulnerabilities and Exposures (CVE) and Common Weakness Enumeration (CWE) data sources and annotated using Wikidata and the SEmantic Processing of Security Event Streams CyberSecurity Knowledge Graph (SEPSES CSKG). Along with the dataset, all the code is publicly released. This dataset is made available to the research community in the context of the SemTab challenge on Tabular to Knowledge Graph Matching. This challenge aims to evaluate the performance of several STI based on open source LLMs. Preliminary evaluation, serving as baseline, was conducted using Falcon3-7b-instruct and Mistral-7B-Instruct, two open source LLMs and GPT-4o mini one closed source LLM.

1 Background & Summary

In the field of cybersecurity, several datasets such as Common Vulnerabilities and Exposures (CVE) [1], Common Attack Pattern Enumeration and Classification (CAPEC) [2], Common Weakness Enumeration (CWE) [3], etc. has been released for various purposes such as development, testing, education, etc. These datasets are used for enabling academics and security professionals to study attack patterns, vulnerabilities, and defence mechanisms; provides data for training and evaluating machine learning (ML) models used in intrusion detection systems, malware analysis, and threat intelligence platforms; offer realistic or synthetic data to test the effectiveness of security tools and techniques [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14].

Security datasets often suffer from limitations. On one hand, Security datasets are scattered on the Internet and provided in heterogeneous formats such as CSV, JSON, XSL, or XML formats. This

makes it difficult to get a holistic view of the interconnectedness of information across different data sources. On the other hand, many datasets focus on specific attack vectors or limited environments, limiting generalisability; There is a lack of detailed annotations in datasets, making it difficult to train supervised learning models. To solve these limits, security data can be extracted from diverse data sources, organised using a tabular data format and linked to existing knowledge graphs (KGs). This is called Semantic Table Interpretation [15, 16]. The KGs schema will help align different terminologies and understand the relationships between concepts.

Although humans can manually annotate tabular data, understanding the semantics of tables and annotating large volumes of data remains complex, resource-heavy and time-consuming [17]. This has led to scientific challenges such as Tabular Data to Knowledge Graph Challenge Matching or SemTab challenge [15, 18, 19]¹. Launched in 2019 and hosted by the International Semantic Web Challenge (ISWC), this challenge aims to benchmark systems dealing with the problem of matching tabular data to KGs. This consists of linking table elements (such as entities in cells, column types, relations between columns) to their corresponding entities in the KG. Although several tabular datasets have been proposed [20, 21, 22, 23], datasets specific in the domain of security are not yet publicly available.

This paper contributes to the SemTab community and the SemTab@ISWC challenge² by introducing the secu-table dataset, composed of tables extracted from CVE and CWE data sources and annotated using Wikidata [24] and SEmantic Processing of Security Event Streams CyberSecurity Knowledge Graph (SEPSES CSKG) [25]. Given that the dataset integrates data extracted from open cybersecurity resources, and knowledge graphs published under the Creative Commons Zero (CC0 1.0 Universal Public Domain Dedication) and the Creative Commons Attribution 4.0 International (CC BY 4.0) licenses, the secu-table dataset released in this work adopted the CC BY 4.0 license to respect the source licenses. The dataset is available at <https://huggingface.co/datasets/jiofidelus/SecuTable>.

The current version of the secu-table dataset (available at https://huggingface.co/datasets/jiofidelus/SecuTable/tree/main/secutable_v2) is composed of more than 1500 tables. These tables contain more than 150k entities, more than 1M lines and more than 20k columns. The average number of columns per table is 8.13 and the average number of rows per table is 291.63. It was made available to the research community in the context of SemTab@ISWC 2025 challenge on Tabular to Knowledge Graph Matching hosted by the 24th International Semantic Web Conference³. It aims at evaluating STI systems, particularly those based on large language models (LLMs).

From January onward, new releases of the dataset will occur on a quarterly basis, with expanded coverage from security data sources such as Adversarial Tactics, Techniques, and Common Knowledge (ATT&CK); Common Configuration Enumeration (CCE); Common Platform Enumeration (CPE); Common Vulnerability Scoring System (CSVS); Open Worldwide Application Security Project (OWASP); Security Content Automation Protocol (SCAP); etc.

2 Methods

This section presents the input data used to create the secu-table dataset and step by step construction method. The first two steps of the method consisted of the recruitment of the data curators and the identification of data sources. Thereafter, the construction pipeline presented by Fig. 1 was executed.

2.1 Input Data

The input data used to construct the current version of the secu-table dataset were derived from the Common Vulnerability and Exposures⁴ (CVE) [1] and Common Weakness Enumeration⁵ (CWE) [3] downloaded in 2022. The dataset obtained was annotated using Wikidata⁶ [24]- a general purpose KG and the SEmantic Processing of Security Event Streams CyberSecurity Knowledge Graph (SEPSES

¹<https://www.cs.ox.ac.uk/isg/challenges/sem-tab/>

²<https://sem-tab-challenge.github.io/2025/>

³<https://sem-tab-challenge.github.io/2025/>

⁴<https://www.cve.org/>

⁵<https://cwe.mitre.org/>

⁶<https://www.wikidata.org/>

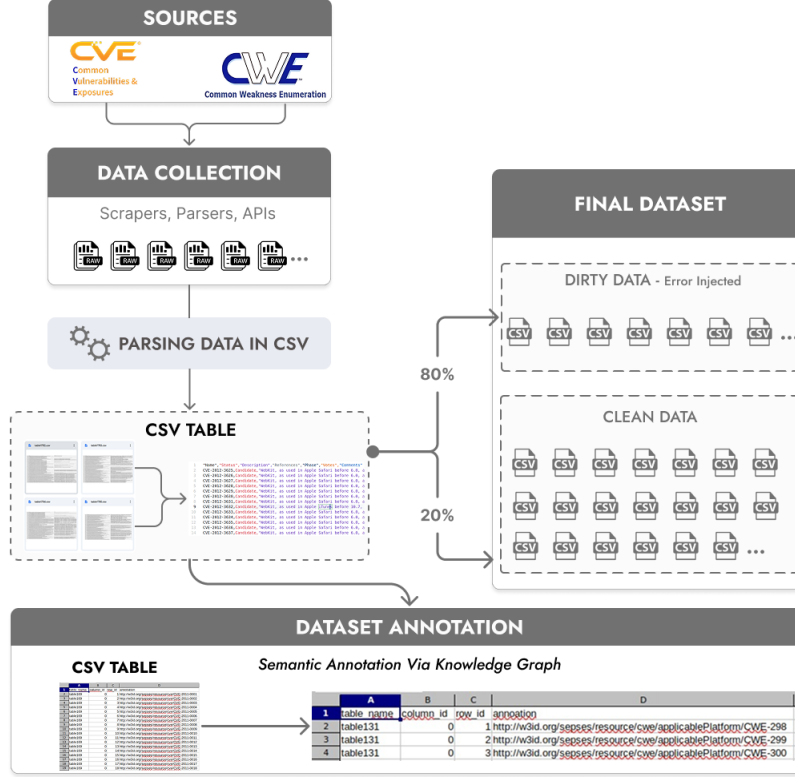


Figure 1: Secu-Table construction pipeline

CSKG) [25]- a security KG. SEPSSES CSKG was chosen because it is the most complete security knowledge graph that is publicly available currently and Wikidata was added to assess the ability of STI systems to cope with more general KG.

2.1.1 Security data sources

CWE [1] is a community-developed list of software and hardware weaknesses. It is widely used in cybersecurity, secure coding, and vulnerability assessment. Each CWE entry describes a common type of mistake or flaw that could lead to exploitable security issues. The CWE list is updated three to four times per year to add new and update existing weakness information. CWE is free to use by any organisation, individual for any research, development, and/or commercial purposes, per the CWE Terms of Use. The version used to build the secu-table dataset (CWE v4.8-June 2022) was downloaded from the CWE repository.

CVE [3] program aims to identify, define, and catalogue publicly disclosed cybersecurity vulnerabilities. It helps track vulnerabilities that need patching, appear in alerts about active exploitation campaigns, used to automate security into continuous integration/continuous delivery pipelines. Each CVE entry contains structured metadata to help identify, track, and remediate the issue. The version used in this work was downloaded in 2022 from CVE (CVE IDs up to CVE-2022-99999) repository.

2.1.2 Knowledge Graphs

A knowledge graph ($KG = (E, R, T)$) is a labelled directed multi-graph in which nodes (E) represent a set of real-world entities, edge types represent relation (R) between nodes and $T \subseteq E \times R \times E$ is a set of triples, where each triple (e_i, r, e_j) represents a directed edge from head entity $e_i \in E$ to tail entity $e_j \in E$ via relation $r \in R$. In this work, we aim to link flat data in security databases to existing KGs so as to add semantics to these tables [26]. To this end, the world largest general purpose KG Wikidata and a security domain specific KG SEPSSES CSKG are used. The Table 1 presents a comparison of these KGs.

Table 1: Comparison of KGs used in this work

KGs	Year	Domain	Model	Entities	Relations	Types
Wikidata	2012	General knowledge	RDF	100M	14B	300K
SEPSES CSKG	2019	Cyber security	RDF	3.8M	479	—

Wikidata [24] is a general purpose, free and open source KG maintained by the Wikimedia Foundation. It aims to store structured data for all sorts of topics, concepts, and objects. Its content is available under a creative commons public domain license. Wikidata helps in knowledge integration by connecting different pieces of information and different datasets, source of standardized identifiers.

The SEPSES CSKG [25] is a cybersecurity KG developed by TU Wien and SBA research group. It integrates and links critical information from publicly available sources. It is constructed using several security data sources such as the Common Weakness Enumeration (CWE) taxonomy, the Common Vulnerabilities and Exposures (CVE) database, the Common Attack Pattern Catalog (CAPEC), and the Security Content Automation Protocol (SCAP). The SEPSES CSKG ⁷ was downloaded and used during table annotation.

2.2 Data curators

Data curators consisted of Master students and one professor in computer science who are co-authors of this paper. These people have a strong background in semantic web, semantic table interpretation and knowledge graphs. The data curators were divided into two groups: the first group consisted of people responsible for the creation of the tabular dataset and the second group of people responsible for the annotation. The most cumbersome and time consuming task was the data annotation. Thus, each data curator was trained on how to find relevant annotations in the different KGs identified. After the training, a qualification test consists of giving five tables to data curators for curation and evaluating their curation by the expert curators. Expert curators have at least two years experience in semantic table interpretation. They were also responsible for quality review, by checking the data curated.

2.3 Data collection & Tables construction

The current version of the secu-table dataset is built using the CWE and CVE entries. These entries involved a unique number, a short description, what the weakness is and why it matters, common consequences, how the weakness might be exploited, real-world cases and the detection and prevention guidance. The data were downloaded from these data sources into different formats such as JSON, XML, CSV. Thereafter, these data were manually parsed to obtain CSV files.

CVE and CWE provide structured data in formats such as JSON, XML, CSV. Therefore, tables were created by exploiting these structured metadata fields. For instance, CWE metadata (e.g., "*CWE – ID*", "*Name*", ... "*Description*", "*RelatedWeaknesses*", ...) extracted from CWE sources were used to construct tables⁸ in which "*CWE – ID*" corresponds to the *ID* of the underlying software weakness, "*name*" corresponds to the weakness name, "*Description*" corresponds to the weakness description, "*RelatedWeaknesses*" corresponds to related weakness, etc. The column content was obtained by grouping entries by considering the CWE metadata under each element (e.g., weakness name under the column "*weakness*").

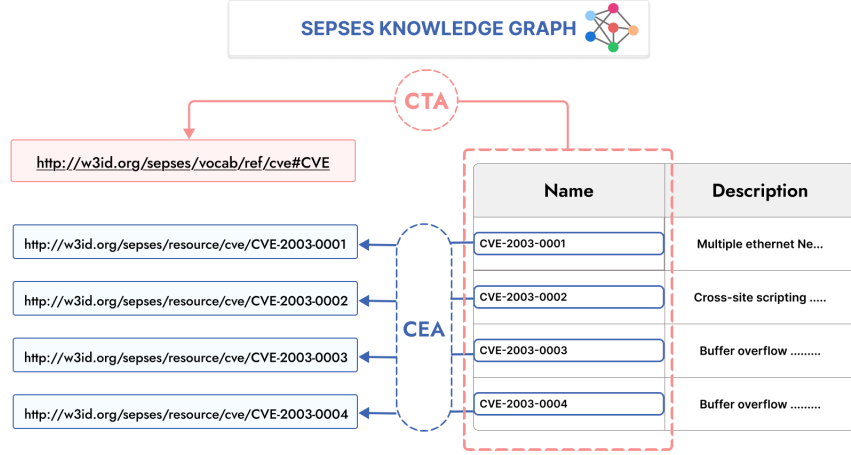
2.4 Dataset annotation

To enhance the understanding of a table, table annotation consists of mapping the table elements to an existing KG. The annotation tasks are illustrated by Fig. 2 and defined by equation 1.

⁷https://sepses.ifs.tuwien.ac.at/dumps/version/102019/graph000001_000001.ttl.gz

⁸https://huggingface.co/datasets/jiofidelus/SecuTable/blob/main/secutable_v2/ground_truth/tables/table1.csv

Figure 2: Illustration of cell entity annotation, column type annotation, and column property annotation using the SEPSES CSKG



- The Cell Entity Annotation (CEA) task consists of mapping table elements to corresponding entities in the KG. Therefore, given an entity e_{tab} in the table tab , the goal is to find its corresponding entity e_{kg} in the KG.
- Column Type Annotation (CTA) task consists of mapping columns in tables to its corresponding types in the KG. Thus, given a column c_{tab} in the table tab , the goal is to find the corresponding type t_{kg} in the KG.
- Column Property Annotation (CPA) task consists of identifying the relation between columns c_{tab_i} and c_{tab_j} in table tab and mapping the latter to corresponding property p_{kg} in the KG.

$$\begin{aligned}
 cea(e_{tab}) &= e_{kg} \\
 cta(c_{tab}) &= t_{kg} \\
 cpa(c_{tab_i}, c_{tab_j}) &= p_{kg}
 \end{aligned} \tag{1}$$

In the equation 1, the cea function takes as input an entity in the table and find its corresponding annotation in the KG. The cta function takes as input a table column (composed of entities stored in this column) and finds the types of the elements contained in this column in the KG and the cpa function takes as input two columns and finds corresponding properties in the KG.

Given that the dataset is being used to evaluate STI during the SemTab challenge in 2025, we need a high quality dataset without errors. Thus, this dataset was manually annotated. It should be noted that manual annotations are generally used to build tabular datasets for evaluating STI [21]. In our case, the annotation is performed with two annotators. During the first round, one annotator provide the different links to the KG. To guarantee the quality of the annotation, a second annotator verifies the data annotated.

Listing 1: SPARQL query allowing to extract CEA annotations for CWE tables. For each CWE with identifier 5, entities and properties are identified from the knowledge graph and extracted

```

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ns2: <http://w3id.org/sepses/resource/cwe/>
SELECT ?cwe ?pre ?obj
WHERE {
  ?cwe a <http://w3id.org/sepses/vocab/ref/cwe#CWE> .
  ?cwe ?pre ?obj .
  FILTER (STRENGTHS(STR(?cwe), "CWE-5"))
}

```

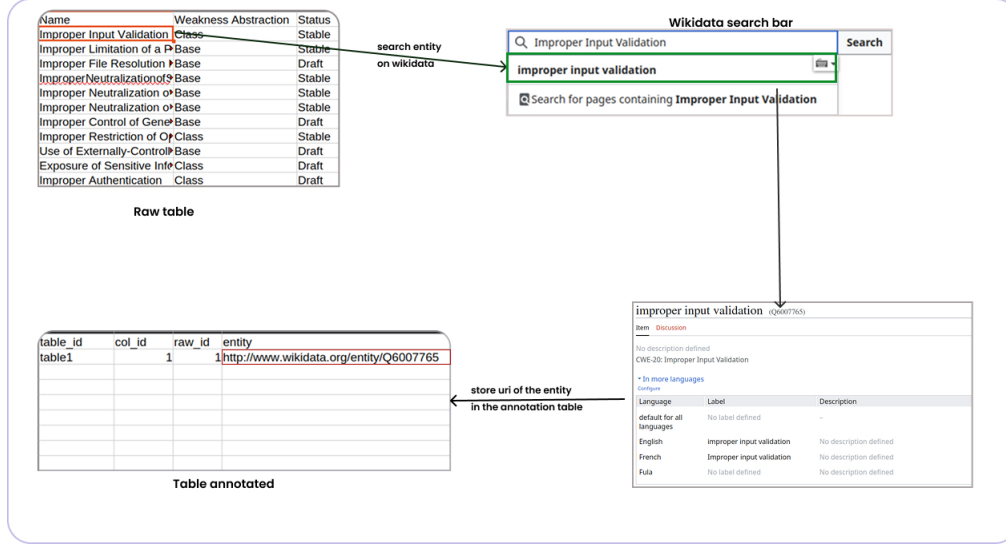


Figure 3: Example of CEA annotation using Wikidata

Listing 2: SPARQL query allowing to extract CEA annotations for CVE tables. For each CVE with identifier 5, entities and properties are identified from the knowledge graph and extracted

```
PREFIX ref: <http://w3id.org/sepses/vocab/ref/cve#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?cve ?property ?value
WHERE {
  ?cve ?property ?value .
  FILTER (CONTAINS(STR(?cve), "CVE-2018-6147"))
}
ORDER BY ?property
```

Annotation using SEPSES CSKG. During SEPSES annotation, we deployed the SEPSES CSKG on a local machine using Jena triple store database⁹. Thereafter, we ran SPARQL queries to search for annotations. The code listing 1 and 2 present the codes that we wrote to retrieve the CEA. After the entities are retrieved, in the case of ambiguity (one query to the KG returns many entities), the contextual information from table rows and columns is leveraged to assess the relevance of candidate annotations and to disambiguate among multiple possible matches.

Annotation using Wikidata KG. Table annotations were identified from Wikidata by manually entering relevant search terms, corresponding to tables elements, entered in the Wikidata search interface (see the illustration of the Fig. 3). Actually, we found that with SPARQL queries, many annotations provided by the query results were not relevant and there were too many ambiguities. Thus, manual search allows us to browse annotations one by one to identify relevant ones. In case of ambiguity (e.g., "Window Server" - corresponds to multiple versions and editions in the KG - the search bar of Wikidata provides "Windows Server 2003 (Q11246)", "Windows Server 2012 (Q11222)", "Windows Server 2008 R2 (Q11226)", etc.), all of which have overlapping labels or aliases. The contextual clues were used from other columns, such as "Version = 2008 R2". Using these clues, the correct entity was selected ("Windows Server 2008 R2 (Q11226)"). We also realised that only a small number of tables were providing annotations. The limited number of available annotations from Wikidata resulted in many empty annotation fields, making it difficult to generate reliable CTA and CPA. Consequently, only the CEA was produced in this version of the dataset.

⁹<https://jena.apache.org/documentation/fuseki2/>

2.5 Dataset construction

The final step of this work consisted of constructing the dataset that will be used for the evaluation of STI. With this dataset, the LLMs should be able to use the table content to detect appropriate annotations. To evaluate the robustness of LLMs for semantic table interpretation, we introduce various types of errors and ambiguities into the dataset using the Pandas library. The dataset was modified as follows:

- 20% data without errors,
- 26% missing context,
- 26.6% misspelling errors,
- and 26.26% annotation errors.

The annotation errors consist of labelling data with wrong labels. These controlled errors simulate real-world challenges in security data annotation, helping to benchmark and improve LLMs performance in handling misspelled, ambiguous and incomplete information.

3 Data record

The secu-table dataset is publicly available for research purposes on Hugging Face¹⁰. Concerning the current release¹¹ (secu-table v2), all data components were downloaded, curated in 2023 for the CEA, CTA and CPA tasks. The dataset contains the following folders:

- **secutable_v2** is the main folder of the dataset, containing all benchmark components necessary for evaluation and replication. This includes the ground truth annotations, raw tables, and test tables.
- **secutable_v2/ground_truth/** contains the ground truth annotations, made available to assist the research community in training, evaluating, and benchmarking STI systems. It includes the following subfolders:
 - **/sepses/**: contains ground truth annotations aligned with the SEPSES CSKG.
 - **/wikidata/**: contains ground truth annotations aligned with Wikidata KG.
 - **/table/**: contains the raw tables that were annotated using both SEPSES CSKG and Wikidata. These tables can be divided into train and test during STI development.
- **/secutable_v2/test/tables/** includes the test tables that researchers can use to evaluate and benchmark their systems. The annotations of these tables are hidden so as to be used as the test for STI systems during the SemTab challenges.

The dataset content indicates that despite manual annotation, several columns still contain empty entries due to the inherent incompleteness of the underlying KG. KG incompleteness is a well-known limitation affecting virtually all existing KGs [26]. KG incompleteness is a common challenge, as many entities and relationships are either partially described or entirely missing in real-world KGs. Consequently, the presence of empty cells reflects this real-world incompleteness and provides a realistic scenario in which downstream systems must handle missing or partial annotations. This design ensures that benchmarking and evaluation capture the practical challenges associated with sparse and incomplete KGs.

This work relies on only two security datasources. In future iterations, we plan to extend the dataset to incorporate additional data sources and knowledge graphs (e.g., DBpedia). Given the high cost and scalability issue of manual annotation, we are exploring a semi-automatic approach combining LLMs (e.g., Falcon3-7b-instruct) with human-in-the-loop to verify the quality of the annotations. Therefore, from January onward, new releases of the dataset will occur on a quarterly basis. These releases will be named according to the modification of the dataset: integration of new tables, integration of new KG or integration of new security data sources.

¹⁰<https://huggingface.co/datasets/jiofidelus/SecuTable/>

¹¹https://huggingface.co/datasets/jiofidelus/SecuTable/blob/main/secutable_v2

4 Data Overview

The current release of the secu-table dataset consists of 1,554 tables, divided into 76 tables provided as ground truth and 1,478 tables for testing and supports the CEA, CTA and CPA annotations. The ground truth table contains more than 8,900 entities, 55,000 rows, 1000 columns. The test tables contain more than 150k entities, 1M rows and 20k columns. The average number of columns per table is 8.13 and the average number of rows per table is 291.63.

5 Technical Validation

The secu-table_v2 dataset was constructed for the purpose of the SemTab@ISWC 2025 challenge. This challenge aims to evaluate the capacity of LLMs to successfully annotate security data using Wikidata and SEPSES KGs. This section presents a preliminary evaluation of the dataset using two open source LLMs and one closed source LLM. The code is provided on GitLab¹² using the MIT license.

Preliminary evaluation consisted of solving the CEA, CTA and CPA tasks on a subset of the dataset composed of 76 tables (ground truth) using Falcon3-7b-instruct¹³ [27], Mistral¹⁴ [28] for the open source LLMs and GPT-4o mini for closed source LLM. LLMs used during evaluation were chosen based on their license (open source), competitive performance against other open and closed source LLMs (using Hugging Face LLM Leaderboard) and their integration with the Hugging Face Transformers library. The following hyperparameters were used for Mistral and Falcon: Temperature=0.7, Top-p=0.95, do-sample=true, use-cache=true, max-new-tokens=512, top-k=50. Given that the task is to assess how well models can annotate, the KG was not explicitly provided to the model as input. A two-shot prompting strategy, where the model receives two illustrative examples in the prompt before the question was used. The baseline, the code and the prompts are provided in the dataset documentation so as to allow newcomers to start from somewhere. The LLMs was run on a 12 core server having the following characteristics: RAM=48Go, GPU type: NVIDIA RTX A3090, GPU memory=48Go.

Semantic table interpretation systems are evaluated based on precision, recall, and F-score [15]. Precision (calculated using 2) is the proportion of correctly predicted annotations out of all predicted annotations. Recall (see equation 3) is the proportion of correctly predicted annotations out of all annotations. And the F1-score (see equation 4).

$$Precision = \frac{\text{relevant annotations}}{\text{total annotations}} \quad (2)$$

$$Recall = \frac{\text{relevant annotations}}{\text{ground truth annotations}} \quad (3)$$

$$F - score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

6 Code Availability

All the source code used during this work is available on GitLab: <https://gitlab.com/fidel.jiomekong/secutable> using the MIT license.

7 Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The work was carried out independently by the authors.

¹²<https://gitlab.com/fidel.jiomekong/secutable>

¹³<https://falconfoundation.ai/>

¹⁴<https://mistral.ai/>

References

- [1] Common Vulnerabilities. Common vulnerabilities and exposures. *The MITRE Corporation*, [online] Available: <https://cve.mitre.org/index.html>, 2005.
- [2] S Barnum. Common attack pattern enumeration and classification (capec) schema. *Department of Homeland Security*, 2008.
- [3] Steve Christey, J Kenderdine, J Mazella, and B Miles. Common weakness enumeration. *Mitre Corporation*, 2013.
- [4] Iram Abrar, Zahrah Ayub, Faheem Masoodi, and Alwi M Bamhdi. A machine learning approach for intrusion detection system on nsl-kdd dataset. In *2020 international conference on smart electronics and communication (ICOSEC)*, pages 919–924. IEEE, 2020.
- [5] Fatima Zohra Belgrana, Nacéra Benamrane, Mohamed Amine Hamaida, Abdellah Mohamed Chaabani, and Abdelmalik Taleb-Ahmed. Network intrusion detection system using neural network and condensed nearest neighbors with selection of nsl-kdd influencing features. In *2020 IEEE International Conference on Internet of Things and Intelligence System (IoTIS)*, pages 23–29. IEEE, 2021.
- [6] Yoshihiro Oyama, Takumi Miyashita, and Hirotaka Kokubo. Identifying useful features for malware detection in the ember dataset. In *2019 seventh international symposium on computing and networking workshops (CANDARW)*, pages 360–366. IEEE, 2019.
- [7] Marian Șandor, Radu Marian Portase, and Adrian Coleșa. Ember feature dataset analysis for malware detection. In *2023 IEEE 19th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 203–210. IEEE, 2023.
- [8] Danish Vasan, Mamoun Alazab, Sobia Wassan, Hamad Naeem, Babak Safaei, and Qin Zheng. Imcfn: Image-based malware classification using fine-tuned convolutional neural network architecture. *Computer Networks*, 171:107138, 2020.
- [9] Tarun Choudhary, Siddhesh Mhapankar, Rohit Bhddha, Ashish Kharuk, and Rohini Patil. A machine learning approach for phishing attack detection. *Journal of artificial intelligence and technology*, 3(3):108–113, 2023.
- [10] Ashit Kumar Dutta. Detecting phishing websites using machine learning technique. *PloS one*, 16(10):e0258361, 2021.
- [11] Felipe Castaño, Eduardo Fidalgo Fernández, Rocío Alaiz-Rodríguez, and Enrique Alegre. Phikita: Phishing kit attacks dataset for phishing websites identification. *IEEE Access*, 11:40779–40789, 2023.
- [12] Fangyi Yu and Miguel Vargas Martin. Honey, i chunked the passwords: Generating semantic honeywords resistant to targeted attacks using pre-trained language models. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 89–108. Springer, 2023.
- [13] Jimmy Dani, Brandon McCulloh, and Nitesh Saxena. When ai defeats password deception! a deep learning framework to distinguish passwords and honeywords. *arXiv preprint arXiv:2407.16964*, 2024.
- [14] Jeremiah Blocki and Peiyuan Liu. Towards a rigorous statistical analysis of empirical password datasets. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 606–625. IEEE, 2023.
- [15] Vasilis Efthymiou, Oktie Hassanzadeh, Mariano Rodriguez-Muro, and Vassilis Christophides. Matching web tables with knowledge base entities: from entity lookups to entity embeddings. In *The Semantic Web—ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21–25, 2017, Proceedings, Part I 16*, pages 260–277. Springer, 2017.
- [16] Azanzi Jiomekong and Brice Foko. Towards an approach based on knowledge graph refinement for tabular data to knowledge graph matching. In *SemTab@ISWC*, 2022.

- [17] Azanzi Jiomekong, Gaoussou Camara, and Maurice Tchuente. Extracting ontological knowledge from java source code using hidden markov models. *Open Computer Science*, 9(1):181–199, 2019.
- [18] Brice Foko, Azanzi Jiomekong, Hippolyte Tapamo, Jérémy Buisson, and Sanju Tiwari. Exploring naive bayes classifiers for tabular data to knowledge graph matching. In *SemTab@ISWC*, pages 72–84, 2023.
- [19] Jean Petit Bikim, Carick Atezong, Azanzi Jiomekong, Allard Oelen, Gollam Rabby, Jennifer D’Souza, and Sören Auer. Leveraging gpt models for semantic table annotation. In *SemTab@ISWC (CEUR Workshop Proceedings, Vol. 3889)*, pages 43–53, 2024, January.
- [20] Madelon Hulsebos, cCaugatay Demiralp, and Paul Groth. Gittables: A large-scale corpus of relational tables. *Proceedings of the ACM on Management of Data*, 1:1 – 17, 2021.
- [21] Nora Abdelmageed, Sirko Schindler, and Birgitta König-Ries. Biodivtab: A table annotation benchmark based on biodiversity research data. In *SemTab@ISWC*, 2021.
- [22] Azanzi Jiomekong, Cosmas Etoga, Brice Foko, Martins Folefac, Sorel Kana, Vadel Tsague, Mouhamadou Sow, and Gaoussou Camara. A large scale corpus of food composition tables. In *SemTab@ISWC*, 2022.
- [23] Vincenzo Cutrona, Federico Bianchi, Ernesto Jiménez-Ruiz, and Matteo Palmonari. Tough tables: Carefully evaluating entity linking for tabular data. In *The Semantic Web – ISWC 2020*, pages 328–343, Cham, 2020. Springer International Publishing.
- [24] Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Communications of the AC*, 57(10):78–85, sep 2014.
- [25] Elmar Kiesling, Andreas Ekelhart, Kabul Kurniawan, and Fajar Ekaputra. *The SEPSES Knowledge Graph: An Integrated Resource for Cybersecurity*, pages 198–214. Springer, 10 2019.
- [26] Philipp Cimiano and Heiko Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semant. Web*, 8(3):489–508, January 2017.
- [27] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023.
- [28] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: [TODO]The main claims made in the abstract and introduction which is the construction of a security tabular dataset for the evaluation of semantic table interpretation systems accurately reflect the paper's contributions and scope. Actually, the dataset is constructed and evaluated (see sections 2, 3, 5).

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: [TODO]The paper discuss the limitations of the work performed (see section 2)

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: [TODO]This is not a theoretical paper

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: [TODO]We explained the dataset construction methodology in section 2 and evaluations, highlighting hyperparameters in section 5

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: [TODO]Hugging Face and GitLab links are provided. These resources are published using the MIT license

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: [TODO]Details are presented in section 5. Additional resources such as annotated data and code description are provided in supplementary materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: [TODO]Experiments results are evaluated with Recall, Precision and F-score metrics

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: [TODO]The details are presented in section 5

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: [TODO]We have read and understood the NeurIPS code of ethics, and have done our best to conform

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: [TODO]This work consists of evaluating the performance of semantic table annotation systems on domain specific datasets and does not impact the society at large.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: [TODO] Since we trained open source LLMs on the dataset built, our work poses no risk of misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: [TODO] We are not shipping the dataset and code from any other existing dataset and code. We did cite open-sourced LLMs and libraries used during this work.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: [TODO]The dataset is released on Hugging Face, with included README files. The README file will be updated with more details documentations

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: [TODO]This work does not involve crowdsourcing nor research with human subjects

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [TODO]This work does not involve crowdsourcing nor research with human subjects

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: **[TODO]**LLMs were used for grammar checking and to polish the quality of the document.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.