From Vector Space to Symbolic Space: Informational and Semantic Analysis of Benign and DDoS IoT Traffic Using LLMs

Mironela Pirnau *, Iustin Priescu *, Mihai-Alexandru Botezatu , Catalina Mihaela Priescu , Daniela Joita

Posted Date: 16 March 2026

doi: 10.20944/preprints202603.1118.v1

Keywords: Large Language Models; security data analysis; semantic encoding

 doi:10.20944/preprints202603.1118.v1

Article

From Vector Space to Symbolic Space: Informational and Semantic Analysis of Benign and DDoS IoT Traffic Using LLMs

Mironela Pirnau [1,*], Iustin Priescu [1,*], Mihai-Alexandru Botezatu [2], Catalina Mihaela Priescu [3] and Daniela Joita [1]

[1] Department of Informatics, Faculty of Informatics, Titu Maiorescu University, 040051 Bucharest, Romania
[2] Department of Informatics, Statistics and Mathematics, School of Computer Science for Business Management, Romanian American University, 012101 Bucharest, Romania
[3] Faculty of Automatic Control and Computer Science, National University of Science and Technology POLITEHNICA Bucharest, Romania
* Correspondence: mironela.pirnau@prof.utm.ro (M.P.); iustin.priescu@prof.utm.ro (I.P.)

Abstract

In this paper, we investigate the feasibility of using Large Language Models (LLMs) for the structural analysis of flow-based network data, considering the fundamental onto-logical difference between the multidimensional numerical space of IoT data and the symbolic space in which these models operate. The primary objective was the development of a formal framework that enables the controlled transformation of numerical data into linguistically analyzable semantic representations, without resorting to classification or machine-learning mechanisms. We propose the SFE mechanism, a deterministic method for robust discretization and behavioral abstraction that converts the numerical characteristics of IoT flows into structural semantic descriptions, based on the CIC IoT-DIAD 2024 [1] dataset. Through formal informational measures, we demonstrate the existence of an intrinsic structural difference between benign and DDoS traffic in the analyzed dataset. In the validation stage, we evaluated whether these informational differences are reflected at the level of linguistic abstraction through controlled inference experiments in IBM WatsonX [2]. The paper demonstrates that LLMs can work as mechanisms for semantic auditing of distributional structure when supported by a formal encoding layer, offering a reproducible framework for integrating numerical security data into language-model-based analysis.

Keywords: Large Language Models; security data analysis; semantic encoding

1. Introduction

Large Language Models (LLMs) [3] are undergoing a rapid process of development and application expansion as they are used not only in natural language processing but also in code analysis, conversational systems, decision support, automated information audits and medical applications such as clinical documentation analysis or diagnostic support. The extension of these models into critical domains has generated increased interest in their use for cybersecurity as well, including log analysis, alert interpretation and network behavior investigation.

The integration of LLMs into security data analysis [4,5] raises a fundamental structural challenge: network data is represented as multidimensional numerical vectors, whereas LLMs operate on symbolic sequences. In the case of flow-based IoT traffic, each instance is described by dozens of continuous statistical features that include communication intensity, temporal structure, and behavioral variability. These representations belong to the $\mathbb{R}^n$ space, while language models process discrete symbolic distributions. The difference between these two spaces is not merely one of format, but of ontological nature. In the absence of a controlled transformation mechanism, the direct

 doi:10.20944/preprints202603.1118.v1

application of LLMs to numerical data risks either the loss of distributional structure or the introduction of artifacts generated by arbitrary conversions.

An important question explored in this paper is whether a deterministic semantic layer can be developed that preserves the informational properties of flow-based data while enabling their coherent analysis in the language space. To answer this question, we propose the Semantic Flow Encoding (SFE) mechanism, a formal transformation from the numerical space into the discrete semantic space. SFE converts numerical vectors of IoT flow into a structured semantic representation through robust discretization based on quintiles and through defining explicit interaction rules between behavioral variables. The role of this mechanism is not classification or predictive performance optimization, but preservation of the distributional structure of the data in a symbolically analyzable form.

Thus, SFE functions as a deterministic intermediate layer between the vector space of numerical features and the language space of LLMs, reflecting in the semantic representation the essential properties of the original distributions. Based on the dataset [1], we have developed distinct semantic corpora for benign traffic and DDoS traffic, and we have evaluated their informational properties prior to any machine learning stage. Shannon entropy analysis allows the measurement of intra-class complexity, while Jensen–Shannon Divergence measures the structural distance between distributions. The results indicate an almost perfect separability in the discrete semantic space, demonstrating that the difference between the two traffic types is an intrinsic property of the distributions rather than an effect of a classification algorithm.

Subsequently, we have investigated whether this informational separability is reflected at the level of linguistic abstraction generated by different foundation models (Granite, LLaMA and Mistral). Through controlled inference in IBM WatsonX, using identical parameters and a single prompt, we observed inter-model semantic convergence in the structural characterization of the corpora. This convergence suggests that the formally demonstrated distributional differences are also detectable in the symbolic space of language representation.

As such, this paper proposes a reproducible framework that rigorously connects the vector space of security data with the symbolic space of language models and positions LLMs as mechanisms for structural auditing of behavioral distributions.

2. Literature Review

The increasingly comprehensive digitalization of modern societies has led to a considerable growth in IoT connections across an expanding range of domains and to the development of business environments based on increasingly distributed infrastructures. These evolutions are accompanied by a series of cyber threats with potentially severe effects on individuals, institutions, and communities. As such, there is a pressing need to ensure secure, fast, and efficient operation and interoperability of IoT systems.

The CIC IoT-DIAD 2024 dataset, designed by the Canadian Institute for Cybersecurity to support cybersecurity researchers, provides a highly diverse and comprehensive collection of data covering a wide range of possible attacks against IoT devices, both for the accurate identification of attacks and for the detection of various forms of operational anomalies in real-world environments.

Recent complex developments have demonstrated that traditional approaches to IoT device identification and anomaly detection in such devices are no longer sufficient and fail to address the full spectrum of attacks that such systems may be exposed to. As such, researchers [1] designed an integrated system featuring modern identification and anomaly detection mechanisms, packet-based feature extraction techniques and functions incorporating specialized attributes for anomaly detection.

To support the development of security analysis applications for IoT devices operating in real-world conditions and to identify solutions to multiple potential attacks, other researchers [6] designed and provided to the research community a realistic dataset with an extended IoT attack

 doi:10.20944/preprints202603.1118.v1

topology—CICIoT2023. This dataset has been employed by a series of researchers to identify new defense mechanisms against continuously evolving cyber threats [7,8].

The large volume and diversity of data in the dataset enable us to identify malicious traffic, attacking devices, and victim devices, contributing to the resolution of security challenges. Similarly, researchers in the healthcare domain [9,10] developed the realistic CICIoMT2024 dataset for the development and evaluation of security solutions in the medical IoMT domain.

The application of LLMs in cybersecurity has increased significantly in recent years. Most existing studies employ LLMs for the analysis and interpretation of textual data, such as logs, alerts generated by IDS/IPS systems or incident reports. In these cases, the models are used primarily as mechanisms for summarization, interpretation, or decision support.

At the same time, the operation of many cybersecurity and IoT systems depends on the characteristics of the underlying network infrastructure. Network latency represents a critical factor for distributed digital applications that rely on real-time communication. Previous studies [11] have shown that variations in network latency can significantly affect the feasibility and reliability of latency-sensitive systems. In parallel, the literature on DDoS attack detection and IoT traffic analysis [12–18] predominantly focuses on machine learning [19] and deep learning models applied directly to numerical flow-based features. Evaluation is performed almost exclusively through classification metrics, and class differences are deduced from the performance of trained models. Although some research uses discretization techniques or feature engineering to improve interpretability, these approaches are oriented toward optimizing predictive performance rather than constructing a formal semantic layer designed for interaction with language models.

The current literature does not explicitly address the problem of controlled transformation of numerical network data into semantic representations that would be usable by LLMs, nor does it systematically investigate the informational separability of classes prior to the machine learning stage. The present paper positions itself within this gap by proposing a deterministic semantic encoding mechanism and by performing a formal, multi-model evaluation of the resulting distributional structure.

3. Dataset Preprocessing and Experimental Framework Foundation

The experimental study utilized data from the publicly available dataset through the Canadian Institute for Cybersecurity (CIC) [1]. From the *Anomaly Detection Flow-Based Features* subset, flows corresponding to benign traffic and those associated with DDoS attacks were selected [20–22]. The initial stage consisted of downloading the files corresponding to benign traffic. The four files associated with normal traffic represent distinct captures recorded in different sessions, sharing a homogeneous structure: the same number of columns (84), identical column order, and the same network flow representation format.

Each file contains 84 flow-based features that numerically describe IoT communication behavior. To obtain a coherent and usable dataset, a preliminary verification of the structural consistency of the four files was performed. This step is essential, as combining files with differing structures could lead to column misalignment or the introduction of missing values. The files were combined vertically, resulting in a consolidated dataset of approximately 398,000 flows with 84 initial features.

The following columns were removed: {Flow ID, Source IP, Destination IP, Source Port, Destination Port, and Timestamp}. These variables describe communication context and infrastructure identifiers rather than the structural behavior of the flow. Their inclusion could have introduced bias or encouraged overfitting to environment-specific identifiers. After removing these six columns, the benign dataset retained 78 numerical features.

Incomplete observations were removed, and invalid values were handled to ensure the numerical integrity of the dataset. Subsequently, a variance analysis was conducted to identify columns with zero variance. Constant variables were eliminated, as they do not contribute to behavioral differentiation. Following this step, the benign dataset was reduced to 70 relevant numerical features. The implicit label *"NeedManualLabel"* was replaced with *"Benign"*.

 doi:10.20944/preprints202603.1118.v1

The same procedure was applied to malicious DDoS traffic. In this case, 14 distinct files were used: 13 corresponding to DDoS-ACK Fragmentation attacks and one associated with DDoS-HTTP Flood [23–25]. After verifying structural consistency (84 initial columns), the files were combined vertically, resulting in a dataset comprising 2,955,806 flows. Removing identifier variables reduced the dataset to 78 features.

During the data cleaning stage [26,27], 2,474 "NaN" values were identified. After removing incomplete observations, the DDoS dataset contained 2,954,569 flows, corresponding to a loss of approximately 0.08% of the total, a statistically negligible impact. Variance analysis identified seven constant columns that were removed, resulting in a final DDoS dataset with 70 numerical features, identical in structure to the benign dataset.

This structural symmetry is fundamental to the comparative validity of the study. Both datasets—benign and DDoS—contain the same number of features, in the same order, and with identical data types. Any differences observed in subsequent analyses can therefore be attributed to behavioral differences rather than structural discrepancies. The underlying assumption of this study is that each instance represents an aggregated IoT flow [1][28], described by statistical behavioral features. The informational analysis conducted in this work characterizes differences in communication behavior between benign and DDoS traffic, rather than differences in content.

4. Semantic Flow Encoding (SFE): Transformation from Vector Space to Symbolic Space

After getting the two clean and structurally symmetric datasets (70 numerical features for *benign traffic* and *DDoS traffic*), the next step consisted of defining a formal mechanism for transforming the numerical representation into a discrete semantic representation compatible with the language analysis performed by LLMs.

Let an IoT flow be represented by a numerical vector: $x = (x_1, x_2, x_3, \dots x_{70}) \in \mathbb{R}^{70}$ where each component represents a continuous flow-based feature. The integration of IoT flow-based data into an analysis framework based on LLMs raises a fundamental structural problem. The analyzed data is numerical, multidimensional, and continuous, organized in a vector space $\mathbb{R}^{70}$, whereas LLMs operate on language sequences belonging to the symbolic space Σ^*. This difference is not merely technical, but ontological: language models process semantic representations, not raw numerical magnitudes.

In our study, each IoT flow is numerically represented by a vector $x \in \mathbb{R}^{70}$ where 70 denotes the number of numerical features remaining after preprocessing [26] (removal of identifiers, missing values, and constant variables). To enable the processing of the files *Benign_All_Flow_Clean_FINAL.csv* and *DDoS_Cleaned.csv* by language models, a controlled transformation from the numerical space to the semantic space was required. For this purpose, the Semantic Flow Encoding (SFE) mechanism was developed. Its role is not classification, but the deterministic transformation of numerical features into a stable semantic representation that can be analyzed from an informational perspective.

The necessity of this intermediate layer derives from the following considerations:

- An IoT flow is described by approximately 70 numerical variables. SFE performs a semantic dimensionality reduction through controlled selection and discretization, selecting a subset of relevant features and reorganizing continuous values into discrete behavioral regimes, without losing the essential behavioral structure.
- IoT data exhibits asymmetric distributions. Robust discretization based on quintiles (20%, 40%, 60%, 80%) allows segmentation of the distribution without assuming normality and reduces sensitivity to outliers.
- Traffic behavior is not determined by isolated variables, but by interactions among them. In the SFE implementation (Appendix 1), explicit behavioral rules are defined within the semantic_encode() function, where ordinal combinations of discretized variables define the final semantic pattern. Listing 1 presents the logical rules used to generate the behavioral label.

 doi:10.20944/preprints202603.1118.v1

Listing 1. Implementation of Behavioral Rules in the semantic_encode() Function

```
# Listing 1: Behavioral interaction rules within semantic_encode()
if throughput in ["high", "very high"] and packet_rate in ["high", "very high"]:
pattern = "burst-like transmission pattern"
elif throughput in ["very low"] and duration in ["high", "very high"]:
pattern = "persistent low-volume communication"
elif variability in ["high", "very high"] and duration in ["very low", "low"]:
pattern = "irregular short burst behavior"
else:
pattern = "stable communication behavior"
```

These rules describe structural relationships between dimensions. A concrete example of the semantic sentence (output of semantic_encode()) for SFE-5 is {A very high TCP-based flow with very low throughput, very low packet rate, very high directional balance, very low packet size variability, exhibiting persistent low-volume communication.}, and for SFE-8 it is {A very high TCP-based flow with very low throughput, very low packet rate, very high directional balance, very low packet size variability, very low average packet size, very high inter-arrival timing, high active phase duration, exhibiting persistent low-volume communication}. It can be observed that, for SFE-8, the sentence becomes longer and more descriptive. It can also be observed that SFE transforms a numerical row from the dataset (70 flow-based columns) into a structured, descriptive natural language sentence, using only robust quintile-based discretization (5 levels: very low / low / moderate / high / very high), as well as simple deterministic rules to determine the dominant behavioral pattern (stable / burst-like / persistent low-volume / irregular short burst), and combination into an intelligible sentence suitable for LLMs. Thus, we obtained a text corpus (one string per flow).

The SFE mechanism is implemented in two configurations: **SFE-5 (k = 5)**, which uses the following fundamental features: Flow Bytes/s; Flow Packets/s; Flow Duration; Packet Length Std; Down/Up Ratio. These features capture traffic intensity, communication density, temporal stability, structural variability and directional balance.

In the SFE-8 configuration (k = 8), three additional features are included: Packet Length Mean (average packet size); Flow IAT Mean (mean inter-arrival time between packets); Active Mean (average duration of active phases). The *k* parameter represents the number of features that were used in the semantic model and defines the semantic dimensionality of the representation: $k \in \{5,8\}$. For each selected feature x_j, the percentile thresholds $Q_{0.2}, Q_{0.4}, Q_{0.6}, Q_{0.8}$ are computed, corresponding to the 20%, 40%, 60%, and 80% percentiles. These thresholds divide distribution into five ordinal intervals: $(-\infty, Q_{0.2}), [Q_{0.2}, Q_{0.4}), [Q_{0.4}, Q_{0.6}), [Q_{0.6}, Q_{0.8}), [Q_{0.8}, \infty)$. The percentile thresholds are separately calculated for each class (Benign and DDoS) in order to preserve internal distributional fidelity and to avoid discretization distortion that would be caused by aggregating heterogeneous distributions. Each interval is mapped to an ordinal semantic label from the set {very low, low, moderate, high, very high}. By applying discretization to the *k* selected features, each flow is represented as $z \in \{1,2,3,4,5\}^k$.

We defined the semantic function as:

$$S: \{1,2,3,4,5\}^k \to \Sigma^* \quad (1)$$

The complete transformation is:

$$F = S \circ D, F: \mathbb{R}^{70} \to \Sigma^* \quad (2)$$

Through this transformation the multidimensional numerical vector becomes a structured semantic sentence, and the entire dataset becomes a semantic corpus. For each class $C \in \{Benign, DDoS\}$, we consider the following set of distinct semantic patterns:

$$S_C = \{s_1, s_2, \dots, s_n\}. \quad (3)$$

The probability distribution is denoted as:

$$P_C(s_i) = \frac{f_i}{N_c}, \quad (4)$$

 doi:10.20944/preprints202603.1118.v1

where f_i represents the frequency of the semantic pattern s_i and Nc is the total number of instances in class C. This distribution describes the behavioral structure of the class in the discrete semantic space. To measure the diversity of each semantic distribution, we compute the Shannon entropy:

$$H(C) = -\sum P_C(s_i)\ log_2\ P_C(s_i)\ . \quad (5)$$

Entropy measures the degree of informational uncertainty associated with a distribution. The interpretation for this is the following: high entropy values indicate a dispersed distribution and high behavioral diversity, while low values indicate probability concentration on a limited number of dominant patterns, suggesting structural rigidity [29–32]. Therefore, $H(Benign)$ and $H(DDoS)$ provide a formal measure of intra-class complexity. In this way we can define the entropy that measures the internal behavioral complexity of each class. To evaluate the difference between the Benign and DDoS distributions, we use the Jensen–Shannon Divergence (JSD), defined as following:

$$JSD(P||Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M) \quad (6)$$

where: $M = \frac{1}{2}(P + Q)$ si D_{KL} represents the divergence Kullback-Leibler [33–35]:

$$D_{KL}(P||Q) = \sum P(i)\ log_2 \frac{P(i)}{Q(i)}. \quad (7)$$

JSD is a symmetric measure, and it is upper-bounded (in base 2) within the interval [0,1]. If $JSD = 0$, then distributions are identical, whereas values close to 1 indicate that the distributions are nearly disjoint. Through this informational analysis, we will simultaneously evaluate intra-class complexity (via entropy) and inter-class structural distance (via JSD). The complete application of the mechanism is presented in Appendix 1 (the semantic_model function and the compute_jsd function), where quintiles are computed, the semantic corpus is generated, Shannon entropy is determined for each class and the Jensen–Shannon Divergence between the semantic distributions is evaluated.

5. Results of the Semantic Informational Analysis

This approach enables the characterization of the difference between benign and malicious traffic at distributional level, independently of any classification algorithm. The semantic layer does not represent merely a textual rephrase of data, but rather a discretized projection that preserves the relevant informational properties of network behavior and allows the formal measurement of class separability.

The results (Tables 1 and 2) clearly highlight that the difference between benign traffic and DDoS traffic is structural and informational in nature, rather than purely numerical. Entropic analysis indicates that the semantic distribution of benign traffic shows higher diversity compared to DDoS traffic in both SFE configurations. In the extended SFE-8 model, the entropy of the Benign class is 7.605218 bits, whereas for DDoS it is 4.889921 bits. The difference of approximately 2.72 bits highlights a reduction in distributional variability in the case of DDoS traffic, suggesting a stronger concentration of probability around a limited number of semantic patterns.

The Jensen–Shannon Divergence reaches 0.999713 in the SFE-8 configuration, approaching the theoretical upper bound (1), while in the SFE-5 configuration the value is 0.967068. These results indicate a pronounced distinction between the semantic distributions of the two classes in the discrete space generated by SFE, reflecting consistent distributional separability between benign and DDoS traffic.

It is important to note that these results are not influenced by a classification mechanism or by any parametric optimization procedure, but they rather derive from the direct application of formally defined informational measures: Shannon entropy, that defines distributional dispersion, and Jensen–Shannon Divergence, that measures the symmetric divergence between two probability distributions. Consequently, the observed difference cannot be attributed to a machine learning stage but reflects a distributional property of the analyzed data. Evaluating informational separability prior

 doi:10.20944/preprints202603.1118.v1

to applying any classification algorithm reduces the risk of interpreting predictive performance as an exclusive effect of the model and contributes to the justification of subsequent conclusions.

The visible structural rigidity in the case of DDoS traffic is compatible with the automated and repetitive nature of distributed attacks, while the semantic diversity of benign traffic is associated with the behavioral variability of legitimate users and applications. This correspondence between mathematical results and behavioral interpretation supports the coherence of the presented conclusions.

The semantic informational analysis indicates that the discretized representation does not introduce evident distortions and allows the highlight of quantifiable differences between classes. The results suggest that the semantic layer can be used as a structural analysis mechanism, providing a formal framework for the subsequent stage, where the generated textual corpora will be utilized within WatsonX platform for further modeling and validation experiments.

Table 1. Results of the SFE-5 model.

Metrics	Benign	DDoS	Notes
Flow numbers	398,198	2,954,569	DDoS has a much larger volume
Distinct semantic patterns	506	263	Benign has greater structural diversity
Uniqueness ratio	0.001271	0.000089	DDoS traffic is significantly more repetitive
Shannon Entropy (bits)	6.517957	4.539669	Benign is informationally more complex
Jensen–Shannon Divergence	0.967068	very high separability	

Table 2. Results of the SFE-8 model (k = 8).

Metrics	Benign	DDoS	Notes
Flow numbers	398,198	2,954,569	DDoS has a much larger volume
Distinct semantic patterns	1,406	664	Benign has greater structural diversity
Uniqueness ratio	0.003531	0.000225	DDoS is significantly more repetitive
Shannon Entropy (bits)	7.605218	4.889921	Benign is informationally more complex
Jensen–Shannon Divergence	0.999713	Very high separability	

Analysis of the Common Semantic Support Between Classes

In parallel with the distributional analysis performed using Shannon entropy and Jensen–Shannon Divergence, we were also interested in identifying the structure of the common semantic support between Benign and DDoS classes. While informational analysis measures the probabilistic difference between distributions, semantic support analysis addresses a more fundamental question: do identical behavioral patterns occur in both classes?

 doi:10.20944/preprints202603.1118.v1

To answer this question, we have implemented the code from Appendix 2, through which we have generated the semantic corpora for each class using the SFE-5 configuration and extracted the sets of distinct patterns. The Python code for Appendix 1 and Appendix 2, together with all data processing files, including the file containing the results obtained using IBM WatsonX (prompt.pdf), are publicly available at the following repository https://github.com/MironelaP/LLM-IoT-DDoS-Analysis. Unlike entropic analysis, where the frequencies of occurrence of each pattern are considered, at this stage we have considered only the existence of patterns, treating each semantic description as an element of a set.

We have computed the intersection of the two sets:

$$S_{common} = S_{Benign} \cap S_{DDoS}$$

This intersection measures the number of semantic patterns that would appear in both classes, independently of their frequency of occurrence. The results are summarized in Table 3.

Table 3. Structure of the common semantic support between classes (SFE-5).

Metrics	Value
Benign semantic patterns	506
DDoS semantic patterns	263
Common semantic patterns	63
Relative overlap in Benign	0.124506
Relative overlap in DDoS	0.239544

The results show that out of the 506 distinct semantic patterns identified in benign traffic, only 63 also appear in DDoS traffic, corresponding to a relative overlap of approximately 12.45%. In the case of DDoS traffic, 63 out of the 263 distinct patterns are shared with Benign, representing approximately 23.95%. These values indicate that there is a limited zone of common semantic behavior between the two classes. Separation is not absolute at the level of discrete support, as certain ordinal combinations of features may appear in both legitimate and malicious traffic. However, the relatively small proportion of shared patterns suggests that most semantic behaviors are specific to 0nly one of the classes.

It is essential to emphasize that this analysis differs conceptually from Jensen–Shannon Divergence [36]. Common semantic support analysis measures the existence of patterns, whereas JSD measures the difference between the probabilistic distributions of patterns. Therefore, the fact that 63 patterns are shared does not contradict the high JSD value. Those patterns may appear in both classes, but with radically different frequencies, which may lead to an almost perfect distributional separability.

From a behavioral perspective the shared patterns may correspond to neutral or ambiguous traffic regimes-for example, short flows with moderate intensity or stable TCP communications-that can occur in both legitimate scenarios and malicious contexts. The major difference between classes does not lie in mere existence of these patterns, but in the way they are distributed.

Therefore, the analysis of common semantic support complements the informational analysis and strengthens the overall conclusion: the difference between Benign and DDoS is not binary at structural level, but it is mostly separable at distributional level. This finding reinforces the robustness of the semantic framework and confirms that the SFE mechanism highlights real and quantifiable differences between the communication behaviors of the two classes.

6. Validation Using Foundation Models in WatsonX

After generating the complete semantic corpora using the proposed Semantic Flow Encoding (SFE) mechanism and demonstrating the informational separability between BENIGN and DDOS traffic through formal measures (Shannon entropy and Jensen–Shannon Divergence), an additional level of interpretative validation was needed. The numerical analysis has confirmed the existence of

 doi:10.20944/preprints202603.1118.v1

a structural difference at distributional level; however, to consolidate the central hypothesis of the paper, it was essential to verify whether this difference would be also detectable at the level of language abstraction, in absence of access to the original numerical values.

For this purpose, a multi-model validation was performed through controlled inference within the IBM WatsonX platform. At this stage, LLMs are used as semantic auditing mechanisms, capable of abstracting the symbolic distribution generated through SFE. The objective is not classification or re-computation of statistical properties, but rather evaluation of semantic structural robustness and the testing of interpretative convergence across distinct models [37–40].

From the completely programmatic generated corpora, representative samples of 1,000 instances were extracted for each semantic configuration:

- BENIGN_SFE-5_1K
- DDoS_SFE-5_1K
- BENIGN_SFE-8_1K
- DDoS_SFE-8_1K

The analysis was organized into BENIGN–DDoS pairs corresponding to each configuration (SFE-5 and SFE-8), keeping both the sample size and the semantic parametrization constant.

The global distributional properties had already been quantified on the complete corpora during the informational analysis stage. The samples preserve the dominant patterns and the relevant informational structure, and processing the entire corpus would not provide significant additional information.

The semantic analysis was conducted in IBM WatsonX Prompt Lab using three distinct foundation models: Mistral, LLaMA-3-70B-Instruct, and Granite (IBM) [41–43]. The choice of the WatsonX platform was motivated by the need for a unified and controlled environment that enables us to run different models with identical generation parameters, ensuring direct comparability of results. For all experiments, identical parameters were used: temperature = 0.2, top-p = 1, frequency penalty = 0, presence penalty = 0, and a constant generation limit (~300 tokens). The low temperature limits stochastic variation and promotes stable inferential behavior. No retrieval mechanism or vector store was used; each file was processed in full as a single context, reducing the risk of introducing selection bias and ensuring a global analysis of the symbolic distribution.

Within this experimental framework, IBM WatsonX provides the necessary infrastructure to test the hypothesis that the informational differences numerically demonstrated between BENIGN and DDOS are systematically reflected in the language abstraction produced by distinct LLMs. The results thus would suggest the existence of a correspondence between the numerical space of distributions and the symbolic space of semantic interpretation. Across all runs, the same prompt was used, formulated in a neutral and descriptive manner, without references to classification or informational measures. The prompt imposed four analytic dimensions: attribute variability, intensity (throughput / packet rate), temporal structure and dominant behavioral motives. Testing was carried out step by step, applying the identical prompt to each pair of datasets and each model, to observe whether the numerically demonstrated informational differences are associated with consistent differences in the generated semantic characterization.

The prompt used was: *You are analyzing a corpus of structured semantic descriptions of network flows. Examine the overall behavioral structure of this corpus. Focus strictly on variability of attributes, intensity patterns (throughput, packet rate), temporal structure (inter-arrival timing, active phase duration) and dominant behavioral motifs. Do NOT classify the data. Do NOT assume labels. Provide a structural analysis of the corpus in approximately 200–300 words.*

The results for each pair of datasets are presented in Tables 4–7. These tables highlight both inter-model convergence and the semantic differences between the BENIGN and DDOS corpora. The comparative tables were produced through a direct and structured analysis of the outputs generated by the three foundation models for each dataset. For each corpus (BENIGN_SFE-5_1K, DDoS_SFE-5_1K, BENIGN_SFE-8_1K, DDoS_SFE-8_1K), the three responses generated by Mistral, LLaMA, and Granite were analyzed separately.

 doi:10.20944/preprints202603.1118.v1

The analysis was organized strictly along the four dimensions that were explicitly imposed by the prompt: attribute variability, intensity (throughput and packet rate), temporal structure, and dominant behavioral motives. For each dimension, the descriptions generated by the three models were compared directly, and the table synthesizes the common ideas and recurrent elements consistently observed across models. No additional automated summarization or classification procedures were applied to the generated outputs. The tables represent a systematic and comparative organization of the content produced by the models, without modification or semantic reinterpretation. The "Convergence" column indicates the degree of conceptual similarity among the descriptions generated by the three models for the same analytical dimension.

The comparative tables synthesize the results generated by each model across four analytical dimensions corresponding to the structure imposed by the prompt. Each row reflects a distinct category of semantic characterization, ensuring that comparisons are performed using homogeneous criteria across models and datasets. The row "Attribute variability" aggregates references to the distribution of descriptive flow values, such as packet size variability, average packet size, or directional balance. This row synthesizes how the models describe the degree of dispersion or amplitude of values along these dimensions, without introducing additional interpretation; it effectively reflects how broad the attribute spectrum is described to be within the analyzed corpus.

The row "Intensity (throughput / packet rate)" consolidates observations related to traffic levels, namely combinations of throughput and packet rate. It synthesizes mentions of very low or very high values, discrepancies between throughput and packet rate, and the way these intensities are distributed across the corpus. The intensity dimension is treated separately from attribute variability to avoid interpretative overlaps.

The row "Temporal" synthesizes elements related to the temporal organization of flows, including explicit references to inter-arrival timing and active phase duration, as well as descriptions of stable or burst-like transmission patterns when correlated with temporal dynamics. This row excludes intensity and focuses exclusively on rhythm, duration, and the temporal distribution of activity.

The row "Dominant motives" reflects how the models articulate recurrent behavioral patterns observed in the corpus, such as stable communication behavior, burst-like transmission, or persistent low-volume communication. This row does not introduce external evaluations but explicitly synthesizes the recurrent narrative structures used by the models to characterize the global behavior of the flows.

The "Convergence" column indicates the degree of semantic similarity among the descriptions generated by the three models for the same analytical dimension. It does not assess performance or accuracy, but rather the consistency of formulations across distinct architectures under identical experimental conditions.

Through this structure the tables do not represent a reinterpretation of the results, but rather a systematic organization of outputs according to the specifications imposed by the prompt, facilitating inter-model and inter-corpus comparison in a transparent and reproducible way.

Table 4. BENIGN_SFE-5_1K.

Dimension	Mistral	Granite	LLaMA	Convergence
Attribute variability	High variability (packet size, directional balance)	Fluctuation de la very low la very high	Variabilities very low → very high	Very strong

 doi:10.20944/preprints202603.1118.v1

Intensity (throughput / rate)	Full spectrum very low → very high	Stable mix + burst-like	Complete range, including mismatch rate/throughput	Strong
Temporal structure	Stable + burst + irregular	Irregular short bursts + persistent low-volume	Stable + burst-like + irregular	Strong
Dominant motives	Stable + burst	Stable + burst	Stable + burst + low-volume	Strong
Global characterization	Divers, heterogeneous	Diverse patterns	Complex and heterogeneous	Very consistent

All three models describe the corpus as heterogeneous, dispersed, and exhibiting wide variation across all dimensions. No clearly dominant core emerges. There is coexistence between stable and burst-like patterns without evident concentration. Inter-model convergence is high.

Table 5. DDOS_SFE-5_1K.

Dimension	Mistral	Granite	LLaMA	Convergence
Variability	High variability, predominantly high directional balance	High packet size variability	Very high packet size variability	Very strong
Intensity	Extreme (very low, very high)	Intense patterns, clear contrasts	Pattern identical to Granite	Very strong
Temporal structure	Burst-like frequently mentioned	Stable + burst, two modes	Stable + burst, two modes	Very strong
Dominant motives	Stable + burst (with emphasis on intensity)	Two modes: steady vs episodic	Two similar modes	Very consistent
Global characterization	More intensity-oriented	Bimodal structure	Bimodal structure	Consistent

Compared with BENIGN_SFE-5_1K (Table 4), in the case of DDOS_SFE-5_1K (Table 5) the models emphasize the recurrent intensity combinations more explicitly as well as the coexistence of distinct behavioral modes, frequently described as steady communication and burst-like transmission. Granite and LLaMA define almost identically the idea of two main behavioral modes: steady transmission and short-term high intensity activities. Even if the variability of the features remains present, the semantic description is better organized around the correlation between throughput, packet rate and temporal patterns associated with these high intensity episodes.

 doi:10.20944/preprints202603.1118.v1

Table 6. BENIGN_SFE-8_1K.

Dimension	Mistral	Granite	LLaMA	Convergence
Variability	Wide variability on multiple attributes	High variability	Large diversity	Very strong
Intensity	Complete spectrum	Complete spectrum	Mix low & high volume	Strong
Temporal structure	Active phase high	Active phase high	Active phase high	Very consistent
Dominant motives	Burst + stable	Burst + stable + low-volume	Burst + stable	Very strong
Global characterization	Diverse	Diverse	Diverse	Very consistent

For **BENIGN_SFE-8_1K** (Table 6), all three models describe the corpus as exhibiting extensive attribute variability, with distributed values across the full spectrum (very low to very high) for throughput, packet rate, and packet size variability. *Active phase duration* is frequently mentioned as being high; however, this feature appears in combination with variable intensities and different temporal patterns (stable and burst-like), without the models explicitly indicating the predominance of a single behavioral type. The overall characterization remains one of structural diversity, convergently supported by all three architectures.

Table 7. DDOS_SFE-8_1K.

Dimension	Mistral	LLaMA	Granite	Convergence
Attribute variability	High variability + directional balance high	High variability	High variability	Very strong
Intensity (throughput / rate)	Mix low & high, sustained	High packet rates frequencies	Low throughput + high-rate pattern	Strong
Temporal structure	Active phase very high	Active phase very high	Active phase very high	Very strong
Dominant motives	Sustained + symmetric	Stable + burst	Stable + burst	Consistent
Global characterization	Dynamic and sustained	Two modes	Burst / continuous	Strong

In the case of **DDOS_SFE-8_1K** (Table 7), all three models frequently show high values of *active phase duration* and elevated levels of *packet size variability* and *directional balance*. Intensity (throughput and packet rate) is described as varying across a wide spectrum, but it is often associated with low inter-arrival timing and episodes of intense activity. Granite and LLaMA formulate very similar descriptions, using comparable conceptual structures to highlight the coexistence of stable and burst-

 doi:10.20944/preprints202603.1118.v1

like behaviors. Semantic characterization is consistent across models and remains stable across all analyzed dimensions.

The comparative analysis of the results for the BENIGN and DDOS datasets highlights a high level of inter-model convergence and consistent semantic differences between the two traffic types. However, although the prompt does not mention classes, the models have independently identified patterns that align with the JSD, and entropy differences shown earlier.

According to Table 4, for the **BENIGN_SFE-5_1K** dataset all three models (Mistral, Granite, and LLaMA) describe the corpus as characterized by extensive attribute variability, a full spectrum of intensity levels (very low to very high), and the coexistence of stable and burst-like patterns without identifying a clearly dominant core. The global characterization is convergent: a diverse, heterogeneous, and structurally dispersed corpus.

In Table 5, corresponding to the **DDOS_SFE-5_1K** dataset, inter-model convergence is again visible, but there are differences in emphasis compared to BENIGN. The models highlight extreme intensities, recurrent combinations of throughput and packet rate, and a clearer structuring of behavior into distinct modes (steady versus burst) more frequently. Granite and LLaMA formulate almost identical descriptions of the existence of two dominant behavioral modes. Compared to Table 4, the structure described is more focused on intensity and recurrent burst-like transmission patterns.

The results for the extended dimensionality SFE-8 maintain this trend. According to Table 6 (**BENIGN_SFE-8_1K**), the benign corpus is again described as exhibiting wide variability across multiple dimensions (*packet size variability*, *average packet size*, *directional balance*), intensities distributed across the full spectrum, and the coexistence of stable and burst-like patterns without evident structural concentration. Inter-model convergence remains strong, and the global characterization is consistent with the one observed for SFE-5.

In Table 7 (**DDOS_SFE-8_1K**), the differences relative to BENIGN become more pronounced. The models frequently emphasize very high *active phase duration*, high *packet size variability*, and intensity-supported patterns. Although formal variability remains present, the descriptions indicate a more coherent structure around burst-like behaviors or sustained high activity. Semantic convergence among Mistral, LLaMA, and Granite is again high, particularly regarding emphasis on intensity and prolonged activity.

Taken together, Tables 4-7 reveal two distinct semantic characterization patterns. BENIGN datasets are consistently described as heterogeneous and dispersed, with broad distributions and no rigid dominance. In contrast, DDOS datasets are repeatedly associated with extreme intensities, bimodal or dual structuring, and the presence of recurrent intense transmission patterns. These differences appear stably regardless of the model used and persist across variations in SFE dimensionality (5 vs. 8).

Therefore, the results suggest robust inter-model convergence and a consistent association between the numerically demonstrated informational differences and the differences observed in language characterizations. Without assuming internal model mechanisms, it can be stated that the generated outputs reflect sensitivity to the distributional structure of the corpora. Thus, the semantic analysis performed through controlled inference in IBM WatsonX provides additional support for the hypothesis that the numerical informational separability between BENIGN and DDOS is associated with stable differences at the level of language abstraction.

7. Theoretical and Practical Conclusions

This paper addressed the problem of the structural difference between the multidimensional numerical representation of flow-based IoT data, and the symbolic representation used by Large Language Models. The main contribution consists in the definition of a formal and deterministic mechanism called **Semantic Flow Encoding (SFE)** which enables the controlled transformation of numerical vectors into a discrete semantic representation that is linguistically analyzable, without resorting to classification or predictive optimization. Informational analysis demonstrated that the difference between BENIGN traffic and DDoS traffic is an intrinsic distributional property, formally

 doi:10.20944/preprints202603.1118.v1

quantifiable through Shannon entropy and Jensen-Shannon Divergence. In both configurations (SFE-5 and SFE-8), the BENIGN class exhibits higher entropy values, indicating greater intra-class structural diversity. In contrast, the semantic distribution of the DDoS class is more concentrated, suggesting a more rigid structuring of behavioral patterns. The high Jensen-Shannon Divergence values (0.967068 and 0.999713) prove pronounced distributional separability in the discrete semantic space.

The analysis of common semantic support complements these results by showing that separation is not absolute at the level of discrete patterns, while major differences emerge at the probabilistic level. This observation confirms the coherence between entropic analysis and support intersection analysis, strengthening the distributional interpretation of the differences between classes.

Validation using foundation models in WatsonX highlights strong inter-model semantic convergence in corpus characterization. The results presented in Tables 4-7 can be directly interpreted in relation to the informational measures reported in Tables 1–3. The higher entropy of the BENIGN class is reflected in convergent linguistic descriptions that characterize the corpora as dispersed and heterogeneous, without a clearly dominant core. By contrast, the distributional concentration of the DDoS class is associated with more structured descriptions, frequently organized around recurrent patterns of intensity and steady or burst-like transmission. The high distributional separability demonstrated numerically is reflected at the language level through consistent differences in emphasis on intensity, sustained activity, and behavioral organization.

Therefore, the semantic analysis performed using foundation models does not introduce artificial differences but reproduces at the language level the numerically quantified structural regularities. The observed convergence across Mistral, LLaMA, and Granite suggests that these characterizations are not artifacts of a particular architecture but reflect stable properties of the semantic distributions generated through SFE.

From a theoretical perspective, this work proposes a reproducible framework that rigorously connects the vector space of security data with the symbolic space of language representation. The contribution does not consist in a new detection model, but in a formal mechanism for transformation and structural analysis, enabling distributional evaluation prior to any machine learning stage.

From an applied perspective, the results indicate that informational separability can be demonstrated before classification, providing an objective foundation for dataset evaluation and reducing the risk of attributing predictive performance exclusively to an algorithm. The SFE semantic layer enables the controlled integration of LLMs into numerical traffic analysis, making it possible to use them as tools for exploration and structural auditing without direct access to raw numerical values. The methodology is extensible to other types of traffic or other flow-based datasets, allowing formal distributional analysis prior to training stages.

Overall, the study shows that the difference between benign traffic and DDoS traffic is detectable and quantifiable at the distributional level prior to classification, and that this difference can be coherently and stably reflected in the symbolic space of language representation. The combination of a formal semantic encoding layer with informational analysis and multi-model validation demonstrates the feasibility of using LLMs as mechanisms for semantic auditing of security data under controlled and reproducible conditions.

The limitations of this analysis stem from the fact that the behavioral rules in the semantic_encode() function are restricted to if-elif conditions and a default case, which simplifies the analysis but may not capture all the complex nuances of feature interactions in highly diverse IoT flows or hybrid attacks. Validation through inference on randomly sampled subsets of 1,000 instances, while sufficient to demonstrate inter-model convergence, does not fully cover the complete corpora consisting of millions of flows, leaving open the possibility of subtle large-scale differences. In future work, we will extend the proposed Semantic Flow Encoding (SFE) mechanism to additional datasets. This extension will allow verification of whether the observed informational differences-higher intra-class entropy for benign traffic versus pronounced concentration for repetitive attacks-

 doi:10.20944/preprints202603.1118.v1

would remain consistent or vary depending on the specific characteristics of each threat type (volumetric, protocol-specific, stealthy, or multi-stage). We also intend to extend semantic validation by testing the SFE mechanism on a broader range of Large Language Models (Gemma, Claude, DeepSeek, etc.), using the same neutral prompt and identical parameters, to confirm the robustness of inter-model convergence and the independence of results from specific model architectures.

Author Contributions: conceptualization, M.P., I.P., M.A.B., D.J.; methodology, C.M.P., and D.J.; software, M.P., C.M.P. and D.J.; validation, I.P., M.A.B. and C.M.P; formal analysis, M.P., I.P., M.A.B, and D.J.; investigation, M.A.B., C.M.P., and D.J.; resources, M.P. and I.P.; data curation, M.P., C.M.P., and D.J.; writing—original draft preparation, I.P., M.A.B., C.M.P., and D.J.; writing—review and editing, M.P, I.P., and D.J.; visualization, M.P., I.P., M.A.B., C.M.P., and D.J.; supervision, I.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data and source code are publicly available at: https://github.com/MironelaP/LLM-IoT-DDoS-Analysis.

Acknowledgments: For the execution of the Python programs and for generating the responses to the prompt used in this study, the authors used the IBM Cloud platform, including IBM Watsonx.ai Studio (Prompt Lab) and Watsonx.ai Runtime services. Access to these resources was made possible through the support provided by the project "Internships in the Academic Center for Artificial Intelligence in the Cloud and with Partners – AI#connect" (Project Code: 312737), implemented under the Education and Employment Program 2021–2027 (PEO) at Titu Maiorescu University in Bucharest, Faculty of Informatics. During the preparation and testing of this study, the authors used IBM watsonx.ai (Prompt Lab) to generate responses to the analyzed prompt. The authors reviewed and analyzed the generated output and took full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
DDoS	Distributed Denial of Service
LLM	Large Language Model
IoT	Internet of Things
DIAD	Device Identification & Anomaly Detection
SFE	Semantic Flow Encoding

References

1. M. Rabbani *et al*., "Device Identification and Anomaly Detection in IoT Environments," in *IEEE Internet of Things Journal*, vol. 12, no. 10, pp. 13625-13643, 15 May15, 2025, doi: 10.1109/JIOT.2024.3522863.
2. IBM WatsonX. Available online: https://www.ibm.com/products/WatsonX (accessed on 27 February 2026).
3. Zhang, L.; Hu, Y.; Li, W.; Bai, Q.; Nand, P. LLM-AIDSim: LLM-Enhanced Agent-Based Influence Diffusion Simulation in Social Networks. Systems 2025, 13, 29. https://doi.org/10.3390/systems13010029
4. Jaffal, N.O.; Alkhanafseh, M.; Mohaisen, D. Large Language Models in Cybersecurity: A Survey of Applications, Vulnerabilities, and Defense Techniques. AI 2025, 6, 216. https://doi.org/10.3390/ai6090216
5. V. Rathod, S. Nabavirazavi, S. Zad and S. S. Iyengar, "Privacy and Security Challenges in Large Language Models," 2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2025, pp. 00746-00752, doi: 10.1109/CCWC62904.2025.10903912.
6. CIC IoT dataset 2023: Neto EC, Dadkhah S, Ferreira R, Zohourian A, Lu R, Ghorbani AA. CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment, Sensors. 2023 Jun 26;23(13): 5941 , https://doi.org/10.3390/s23135941

 doi:10.20944/preprints202603.1118.v1

7. Jony, A.I., & Arnob, A.K. (2024), A long short-term memory based approach for detecting cyber attacks in IoT using CIC-IoT2023 dataset, Journal of Edge Computing, 2024, 3, 28-42, https://doi.org/10.55056/jec.648
8. Thaer AL IBAISI (2025). CICIoT2023 Dataset. IEEE Dataport. https://dx.doi.org/10.21227/v63c-9998
9. Sajjad Dadkhah, Euclides Carlos Pinto Neto, Raphael Ferreira, Reginald Chukwuka Molokwu, Somayeh Sadeghi, Ali A. Ghorbani, CICIoMT2024: A benchmark dataset for multi-protocol security assessment in IoMT, Internet of Things, Vol 28, 2024, 101351, ISSN 2542-6605, https://doi.org/10.1016/j.iot.2024.101351
10. Sajjad Dadkhah, Euclides Carlos Pinto Neto, Raphael Ferreira, Reginald Molokwu, CICIoMT2024: Attack Vectors in Healthcare devices-A Multi-Protocol Dataset for Assessing IoMT Device Security, February 2024, Preprints.org, DOI:10.20944/preprints202402.0898.v1
11. K. Orwa, Y. Chen, V. Bathija and M. Teodorescu, "Real-Time Surgery, Delayed: Internet Latency and the Prospect of Telerobotic Surgery," 2025 IEEE Global Humanitarian Technology Conference (GHTC), Golden, CO, USA, 2025, pp. 1-8, doi: 10.1109/GHTC66843.2025.11266771.
12. Wahab SA, Sultana S, Tariq N, Mujahid M, Khan JA, Mylonas A., A Multi-Class Intrusion Detection System for DDoS Attacks in IoT Networks Using Deep Learning and Transformers, Sensors (Basel), 2025 Aug 6; 25(15):4845. doi: 10.3390/s25154845, PMID: 40808008; PMCID: PMC12349258, https://pmc.ncbi.nlm.nih.gov/articles/PMC12349258/
13. Alshdadi, A. A., Almazroi, A. A., Ayub, N., Lytras, M. D., Alsolami, E., & Alsubaei, F. S. (2024). Big Data-Driven Deep Learning Ensembler for DDoS Attack Detection, Future Internet, 16(12), 458. https://doi.org/10.3390/fi16120458
14. A. Ramzy Shaaban, E. Abdelwaness and M. Hussein, "TCP and HTTP Flood DDOS Attack Analysis and Detection for space ground Network, "2019 IEEE International Conference on Vehicular Electronics and Safety (ICVES), Cairo, Egypt, 2019, pp. 1-6, DOI: 10.1109/ICVES.2019.8906302
15. Indraneel Sreeram, Venkata Praveen Kumar Vuppala, HTTP flood attack detection in application layer using machine learning metrics and bio inspired bat algorithm, Applied Computing and Informatics, Volume 15, Issue 1, 2019, Pages 59-66, ISSN 2210-8327, https://doi.org/10.1016/j.aci.2017.10.003 .
16. IBM, What is a distributed denial-of-service (DDoS) attack?, Jim Holdsworth, Matthew Kosinski, https://www.ibm.com/think/topics/ddos#1743856945
17. R. Sanjeetha, K. N. A. Shastry, H. R. Chetan and A. Kanavalli, "Mitigating HTTP GET FLOOD DDoS attack using an SDN controller,"2020 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), Bangalore, India, 2020, pp. 6-10, doi: 10.1109/RTEICT49044.2020.9315608 .
18. Karanpreet Singh, Paramvir Singh, Krishan Kumar, Application layer HTTP-GET flood DDoS attacks: Research landscape and challenges, Computers & Security, Vol 65, 2017, Pages 344-372, ISSN 0167-4048, https://doi.org/10.1016/j.cose.2016.10.005
19. Abdullah Alabdulatif, Navod Neranjan Thilakarathne, Mohamed Aashiq, Machine Learning Enabled Novel Real-Time IoT Targeted DoS/DDoS Cyber Attack Detection System, Computers, Materials and Continua, 2024, Vol 80, Issue 3, 2024, Pag 3655-3683, ISSN 1546-2218, https://doi.org/10.32604/cmc.2024.054610 .
20. Alashhab, Z.R.; Anbar, M.; Rihan, S.D.A.; Alabsi, B.A.; Ateeq, K. Enhancing Cloud Computing Analysis: A CCE-Based HTTP-GET Log Dataset. Appl. Sci. 2023, 13, 9086. https://doi.org/10.3390/app13169086
21. Mahjabin, Tasnuva, et al. "Load distributed and benign-bot mitigation methods for IoT DNS flood attacks." IEEE Internet of Things Journal 7.2 (2019): 986-1000.
22. Li, Q.; Liu, Y.; Niu, T.; Wang, X. Improved Resnet Model Based on Positive Traffic Flow for IoT Anomalous Traffic Detection. Electronics 2023, 12, 3830. https://doi.org/10.3390/electronics12183830
23. Pakmehr, A., Aßmuth, A., Taheri, N. et al., DDoS attack detection techniques in IoT networks: a survey, Cluster Computing 27, 14637–14668 (2024), https://doi.org/10.1007/s10586-024-04662-6
24. Al-Hadhrami, Y., Hussain, F.K. DDoS attacks in IoT networks: a comprehensive systematic literature review, World Wide Web 24, 971–1001 (2021). https://doi.org/10.1007/s11280-020-00855-2
25. Vishwakarma, R., Jain, A.K. A survey of DDoS attacking techniques and defence mechanisms in the IoT network, Telecommun Syst , 73, 3–25 (2020). https://doi.org/10.1007/s11235-019-00599-z

 doi:10.20944/preprints202603.1118.v1

26. M.H. Teodorescu, “Natural language processing techniques in management research ”. Chapter 3 in Research Handbook on Artificial Intelligence and Decision Making in Organizations, Eds. I. Constantiou, M. Joshi, and M. Stelmaszak, pp. 58–79. Edward Elgar Publishing, Cheltenham, 2024, UK. DOI: https://doi.org/10.4337/9781803926216
27. Zaoui Seghroucheni, O.; Lazaar, M.; Al Achhab, M. Using AI and NLP for Tacit Knowledge Conversion in Knowledge Management Systems: A Comparative Analysis. Technologies 2025, 13, 87. https://doi.org/10.3390/technologies13020087
28. Mutambik, I. An Efficient Flow-Based Anomaly Detection System for Enhanced Security in IoT Networks. Sensors 2024, 24, 7408. https://doi.org/10.3390/s24227408
29. C. Bolea, M.H. Teodorescu, S. Bejinariu, D. Gifu, H.N. Teodorescu, V. Apopei, “Similarity Computation based on the Tails of the Rank Distributions and the Related Graphs,” IEEE High Performance Extreme Computing Conference Proceedings 2023
30. Nielsen, F. On a Generalization of the Jensen–Shannon Divergence and the Jensen–Shannon Centroid. Entropy 2020, 22, 221. https://doi.org/10.3390/e22020221
31. Nielsen F., On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means. Entropy. 2019; 21(5):485. https://doi.org/10.3390/e21050485
32. Miranda, F.; Balbi, P.P. Simulating Public Opinion: Comparing Distributional and Individual-Level Predictions from LLMs and Random Forests. Entropy 2025, 27, 923. https://doi.org/10.3390/e27090923
33. Pistone, G. Affine Calculus for Constrained Minima of the Kullback–Leibler Divergence. Stats 2025, 8, 25, https://doi.org/10.3390/stats8020025
34. Amari, S.I. Information Geometry and Its Applications; Applied Mathematical Sciences; Springer: Tokyo, Japan, 2016; Volume 194, pp. xiii+374. [Google Scholar]
35. Lang, S. Differential and Riemannian Manifolds, 3rd ed.; Graduate Texts in Mathematics; Springer: Berlin/Heidelberg, Germany, 1995; Volume 160, pp. xiv+364. [Google Scholar]
36. Lionis, A.; Peppas, K.P.; Nistazakis, H.E.; Tsigopoulos, A. RSSI Probability Density Functions Comparison Using Jensen-Shannon Divergence and Pearson Distribution. Technologies 2021, 9, 26. https://doi.org/10.3390/technologies9020026
37. M. Son and S. Lee, "Performance Analysis of Prompt-Engineering Techniques for Large Language Model," 2025 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 2025, pp. 1-5, doi: 10.1109/ICCE63647.2025.10930066.
38. A. Rula, J. D'Souza, “Procedural Text Mining with Large Language Models,” K-CAP '23: Proceedings of the 12th Knowledge Capture Conference 2023, Pensacola FL USA December 5–7, 2023, pp. 9–16, https://doi.org/10.1145/3587259.362757
39. Meta AI, “Llama 2: Open Foundation and Fine-Tuned Chat Models,“ArXiv, https://doi.org/10.48550/arXiv.2307.09288 .
40. Vivian Liu and Lydia B. Chilton. Design guidelines for prompt engineering text-to-image generative models, 2023.
41. Son, M.; Won, Y.-J.; Lee, S. Optimizing Large Language Models: A Deep Dive into Effective Prompt Engineering Techniques. Appl. Sci. 2025, 15, 1430. https://doi.org/10.3390/app15031430
42. Gershon, Talia, et al. "The infrastructure powering IBM's Gen AI model development." arXiv preprint arXiv:2407.05467 (2024).
43. Cruz, Rogelio, et al. "Prompt engineering and framework: implementation to increase code reliability-based guideline for LLMs." arXiv preprint arXiv:2506.10989 (2025).