Article

Not peer-reviewed version

Generative Artificial Intelligence and Probabilistic Trees for the Linguistic Data Summarization in Wave Energy Decision-Making

Iliana Pérez Pupo* *, Luis Alvarado Acuña , Pedro Y. Piñero Pérez , Raykenler Yzquierdo Herrera , Maikel Yelandi Leyva Vázquez

Posted Date: 20 April 2026

doi: 10.20944/preprints202604.1304.v1

Keywords: linguistic data summarization; Chow–Liu algorithm; uncertainty; artificial intelligence; decision-making; oceanography

Article

Generative Artificial Intelligence and Probabilistic Trees for the Linguistic Data Summarization in Wave Energy Decision-Making

Iliana Pérez Pupo*[1,2], Luis Alvarado Acuña[1,2], Pedro Y. Piñero Pérez[1,2,3], Raykenler Yzquierdo Herrera[3] and Maikel Yelandi Leyva Vázquez[4]

[1] IADES ALLIANCE, International Alliance for Artificial Intelligence for Sustainable Development, Chile
[2] IADES, Sociedad Mercantil
[3] Universidad Católica del Norte, https://ror.org/02akpm128, Consorcio HEUMA, Facultad de Ciencia de la Ingeniería en la Construcción, Departamento de Gestión de la Construcción, Chile
[4] Centro de Investigación Desarrollo e Innovación en Inteligencia Artificial, Universidad Central del Este, Dominican Republic
[5] Facultad de Ciencias Matemáticas y Físicas, Universidad de Guayaquil, Ecuador
* Correspondence: iliana.perez@ce.ucn.cl; Tel.: +56 9 7210 9615

Abstract

This paper presents a hybrid model that combines linguistic data summarization techniques, algorithms for constructing probabilistic trees, and various Generative Artificial Intelligence models for learning and generating linguistic summaries to aid decision-making. The proposal is validated using methodological triangulation techniques that demonstrate high consistency in the knowledge discovered. The proposal also compares different Generative Artificial Intelligence models; among the evaluated models, Gemini achieved the best performance. However, it is evident that in certain contexts and tasks, small language models can be effective, yielding results comparable to Large Language Models (LLMs) at a lower computational cost. This study applies the algorithms in a case study analyzing oceanographic data from northern Chile. In the validation scenario, the combination of linguistic data summarization methods with unsupervised learning techniques effectively models human tolerance for imprecision when processing complex data and generated linguistic summaries easily interpretable by human decision-makers with high levels of confidence. Studies of energy capacities in the studied region and their behavior in both winter and summer are presented.

Keywords: linguistic data summarization; Chow–Liu algorithm; uncertainty; artificial intelligence; decision-making; oceanography

---

1. Introduction

Decision-making in complex domains is increasingly challenged by growing data volumes and inherent uncertainty. Modern information systems generate vast amounts of numerical and categorical data. However, these representations are often difficult for humans to interpret, leading to 'data-rich but information-poor' environments [1]. In this context, methods capable of transforming raw data into interpretable, reliable, and uncertainty-aware knowledge representations are essential.

Linguistic Data Summarization (LDS) has emerged as a prominent descriptive knowledge discovery technique. LDS enables the generation of summaries that are both semantically meaningful and cognitively accessible, facilitating decision-making even for non-expert users. This approach has been extensively studied within the fuzzy logic and computational intelligence communities, where protoforms and linguistic quantifiers serve as the core mechanism for representing generalized data behavior [2,3,4,5,6].

Despite its interpretability, LDS struggles to automatically identify relevant at-tribute combinations and meaningful relationships, especially in high-dimensional datasets. Exhaustive search strategies are computationally infeasible and often produce trivial or statistically weak summaries. Moreover, many LDS approaches lack explicit mechanisms to control uncertainty. As a result, they are vulnerable to spurious associations caused by noise, sampling variability, or co-incidental correlations.

Probabilistic graphical models provide a principled framework for addressing these limitations by explicitly modeling dependencies among variables. Among them, the Chow–Liu algorithm offers an efficient method for learning dependency trees based on mutual information, yielding a compact approximation of multivariate probability distributions. Its ability to extract strong pairwise dependencies make it suitable as an intermediate layer to reduce the search space and guide the construction of candidate summaries [7]. However, dependency trees produce mathematically sound representations that are not inherently human-interpretable. This limits their direct use in decision-making. Thereby motivating the use of linguistic data summarization and Generative Artificial Intelligence (GenAI, for brevity) to translate these outputs into natural, suitable for decision-making.

When combined with structured and validated knowledge representations, generative models can produce linguistic summaries that are coherent, traceable, and explainable. In this framework, LDS artifacts act as an intermediate, human-consistent knowledge layer that can be injected into generative models through mechanisms such as retrieval-augmented generation and modular knowledge adapters, reinforcing decision-making processes under uncertainty.

This paper presents a framework that integrates the Chow–Liu algorithm [8], statistical validation, and GenAI for uncertainty-aware linguistic data summarization. The approach is validated through a real-world oceanographic case study. Results show improved fidelity, interpretability, and reliability of the summaries. By combining probabilistic modeling and GenAI, the method supports robust decision-making in uncertain and data-intensive environments.

The remainder of the article is organized as follows. Section 2 presents an analysis of the use of probabilistic networks in the linguistic summarization of data, situating the proposal within the state of the art and highlighting the limitations of existing approaches. Section 3 describes the proposed algorithmic framework, based on the Chow-Liu algorithm and statistical validation techniques, as well as its integration with GenAI models for generating interpretable linguistic summaries. Section 4 presents the experimental evaluation through a case study with real oceanographic data, along with a comparative discussion of the results obtained with different generative models. Finally, Section 5 summarizes the main conclusions and proposes avenues for future research.

2. Materials and Methods

2.1. *Study of the State of the Art of Probabilistic Models and Linguistic Summarization of Data*

Since 2021, research in Linguistic Data Summarization (LDS) has experienced sustained growth, expanding both its theoretical foundation and its areas of ap-plication. Recent advances reflect an evolution from classical approaches based on fuzzy logic toward more integrated proposals that incorporate deep learning, explainable artificial intelligence (XAI) techniques, advanced temporal modeling, and hybrid methods for knowledge discovery. Among these contributions are the use of fuzzy linguistic summaries to explain unsupervised online learning processes, the development of temporal protoforms capable of describing com-plex dynamics [1, 9, 10].

Furthermore, unsupervised and active learning approaches have been explored, along with hybrid methods based on approximate and fuzzy sets, with the aim of improving the descriptive capacity and semantic relevance of the generated summaries. These lines of work demonstrate a growing interest in providing LDS with greater robustness, adaptability, and explanatory power, especially in contexts characterized by large data volumes and high uncertainty [1]. This study

 doi:10.20944/preprints202604.1304.v1

highlights the potential of probabilistic methods for information synthesis, constituting a relevant precedent for this research.

However, a systematic review of the literature indexed in the main academic search engines, specifically Web of Science (WoS) and Scopus, reveals that the explicit use of probabilistic networks for generating linguistic summaries of data remains limited, as shown in Fig. 1. This low level of adoption becomes even more evident when analyzing its evolution over time (see Fig. 2), where a sporadic and unconsolidated presence of this type of approach is observed. Even within the classic lines of research promoted by Yager, Kacprzyk, and Zadeh, research has predominantly favored fuzzy mechanisms and linguistic formulations, relegating the systematic exploitation of graphical probabilistic models as structural tools for constructing linguistic summaries to a secondary role.

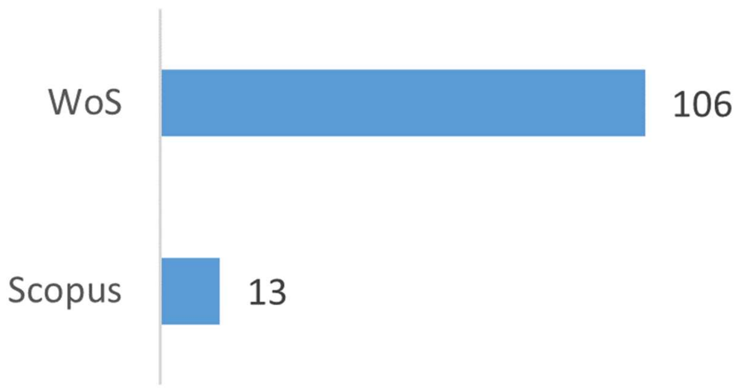
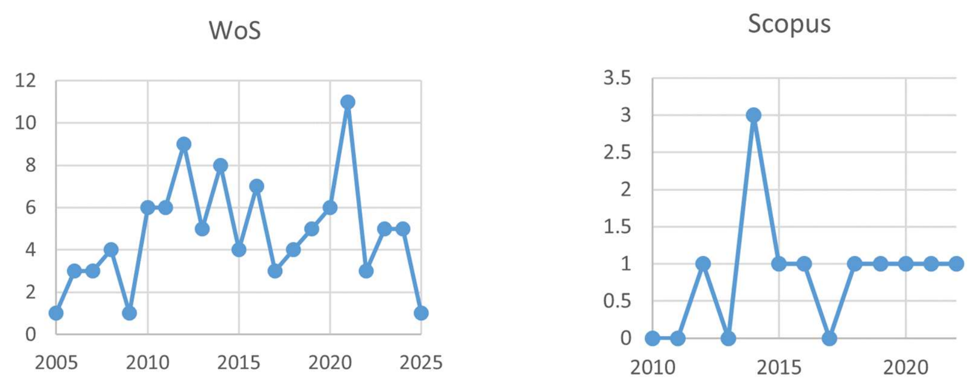


Figure 1. Publications on probabilistic networks in Linguistic Data Summarization (WoS and Scopus).



(**a**) (**b**)

Figure 2. Temporal evolution on probabilistic networks in Linguistic Data Summarization (WoS vs Scopus). (**a**) Temporal evolution in WoS; (**b**) Temporal evolution in Scopus.

In contrast, probabilistic networks have been widely used in other areas of natural language processing (NLP), as evidenced by Fig. 3, where a sustained and clearly growing scientific output in Web of Science and Scopus is observed, associated with the combination of probabilistic models and NLP techniques.

 doi:10.20944/preprints202604.1304.v1

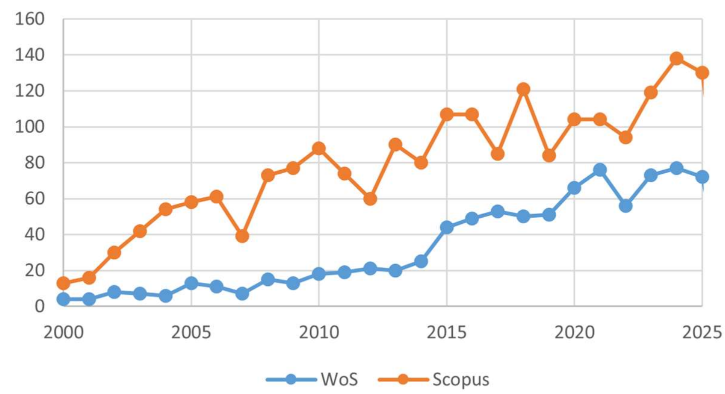


Figure 3. Publication trends on probabilistic approaches in Natural Language Processing (WoS and Scopus).

Unlike the specific context of linguistic data summarization, where the use of probabilistic networks remains limited and fragmented, in NLP these models have demonstrated their ability to capture complex structural dependencies, handle the uncertainty inherent in natural language, and improve the coherence and interpretability of the results. This gap highlights a clear opportunity for methodological transfer: the probabilistic principles that have proven effective in NLP can be adapted and reinterpreted to strengthen the generation of linguistic data summaries, particularly as intermediate mechanisms for structuring knowledge before its verbalization through GenAI.

The work closest to the use of probabilistic graphs in this field is Pupo's proposal [7], which employs the LPA algorithm to generate sets of probabilistic trees from which candidate summaries are constructed. This algorithm, with a time complexity on the order of $O(n^3)$, allows the identification of the strongest relationships between variables through probabilistic learning and the production of linguistic summaries that reduce the search space and facilitate the processing of heterogeneous data. This approach introduces a key idea: using probabilistic structures as a preliminary mechanism for filtering and organizing knowledge before its linguistic expression.

The algorithm consists of two main stages [7]. The first stage is associated with discovering the intrinsic relationships in the data. The second stage is associated with generating summaries from the learned relationships. Generating these summaries involves describing controlled natural language grammars, linguistic variables with fuzzy sets for each variable, and dictionaries for translation. One of the limitations of this algorithm is that it requires a large amount of data to generate natural language expressions, which hinders its applicability to real-world problems.

A solution to this complex problem can be found in leveraging the capabilities of GenAI and its potential for natural language processing. GenAI capabilities, combined with probabilistic tree learning, provide a powerful tool for modeling relationships between variables, overcoming the limitations of previous algorithms.

The next section presents a proposal that improves upon the shortcomings of the LPA_LDS algorithm proposed in [7].

2.2. *Proposed Method*

New Method for Linguistic Data Summarization Combining GenAI and Probabilistic Trees

This section presents the proposed hybrid framework for Linguistic Data Summarization (LDS), which integrates probabilistic graphical models with GenAI to produce interpretable and uncertainty-aware linguistic summaries.

The method is designed to address two key limitations of traditional LDS approaches: (i) the combinatorial complexity associated with exhaustive search of attribute relationships, and (ii) the lack of explicit mechanisms to model and control uncertainty in the generated summaries. To overcome these limitations, the proposed approach introduces a statistical–generative pipeline that combines probabilistic structure learning with controlled natural language generation.

 doi:10.20944/preprints202604.1304.v1

The framework operates in three main stages:

- First, a probabilistic model is learned from the input dataset using graph-based learning techniques. In this work, the Learning Polytree Algorithm (LPA) is employed to identify the strongest dependencies between variables and to construct a set of probabilistic trees that approximate the joint distribution of the data.
- Second, candidate linguistic summaries are generated from the structure of the probabilistic trees. Each tree defines a set of relationships where the root node is interpreted as the summarizing attribute and the connected nodes as filtering conditions. This transformation reduces the search space and ensures that only statistically relevant relationships are considered.
- Third, GenAI models are used to transform candidate summaries into natural language expressions. This process is guided by a controlled natural language (CNL) grammar based on linguistic protoforms, ensuring that the generated summaries are semantically consistent, interpretable, and aligned with human reasoning patterns.

The overall workflow of the method integrates statistical learning and language generation into a unified pipeline, enabling the construction of linguistic summaries that are both mathematically grounded and cognitively meaningful. Additionally, the framework incorporates quality evaluation metrics (T1–T7) to assess the validity, precision, coverage, and interpretability of the generated summaries.

This hybrid approach allows the system to leverage the strengths of probabilistic modeling for knowledge extraction and GenAI for natural language production, resulting in a scalable and robust solution for decision-making under uncertainty.

2.3. Algorithm Description

The proposed algorithm uses probabilistic graph learning techniques to discover relationships between variables [11].

The proposed algorithm does not include an explicit data preprocessing stage, as it assumes that the input data have been previously processed. Thus, standardized data quality is a prerequisite for the execution of the model.

From the graph learning, a set of probabilistic trees is generated that represent the different relationships between the variables (see algorithm 1).

Algorithm 1 Generation of linguistic summaries combining probabilistic models and GenAI

Input

Dataset $U = \{x_1, x_2, ..., x_n\}$, where each x_i is an object (row) described by p attributes, and $n=|U|$ is the total number of objects.

Controlled Natural Language definition LNC=($LNC_{Grammar}$), where $LNC_{Grammar}$ denotes the grammar of the linguistic summaries, contains simple phrases that describe the variables and attributes of the problem.

List of indicators $T = (T_1, T_2, ..., T_7)$ to quality assessment of linguistic summaries.

Output:

Set of linguistic summaries S

Quality evaluation scores $Q(S)$

Begin

Candidates = ∅

Step 1. *Clean data algorithms*

Step 2. *prob_graph = build_probabilistic_graph(U)*

Step 3. *For each of the prob_graph $_i$ trees in prob_graph do*

Step 4.1. *If prob_graph $_i$ has more than one vertex*

Step 4.2. *candidates$_i$= do_candidate_from_branches (prob_graph $_i$)*

Step 4.3. *Candidates ← candidate_summaries$_i$*

End of the conditional statement started in step 3.1

End of the cycle started in step 3

Step 4. *Summaries = agent_generate_summaries(U, Candidates, LNC)*

Step 5. *calculate_T(U, Summaries, T)*

Step 6. *return Summaries*

End

Essentially, the algorithm in step 2 transforms the input data by performing a data cleaning process. This involves applying standardization techniques, removing incomplete data, and detecting and eliminating outliers.

Then, in step 3, an algorithm is applied to the transformed dataset to learn the probabilistic model [12] that best approximates the behavior of the data. Examples of algorithms that can be used in this step are listed below:

- The Chow Liu algorithm, developed in 1968, was initially proposed for constructing trees from data [11].
- The Rebane-Pearl algorithm [13] extends the Chow Liu algorithm and enables the learning of polytrees, allowing the description of higher-order interactions [11].
- The Polytree Approximation Algorithm (PA), developed in 1998, bases its learning on the calculation of marginal and conditional mutual information.
- The Learning Polytree Algorithm (LPA) is the algorithm used in the experiment and is recommended in this research. It was proposed in [14], and its main contribution is improved efficiency compared to previous algorithms [12]. This algorithm calculates marginal and conditional dependencies using computational methods based on the concept of entropy. The complexity of this algorithm is $O(n^3)$ [12].

The algorithms in step 3, which learn the probabilistic models, generate a polytree where each tree can generate a candidate summary. Then, in step 4, a candidate object is generated for each tree, such that the node at the root of the tree is identified as the summarizing attribute and the nodes on the branches are identified as filters. The number of trees that can be generated is bounded above by the total number of attributes, p; the maximum number of elements in filters and summarizing objects is p-1. Therefore, the complexity of this step is bounded by p(p - 1), which is $O(p^2)$.

In step 5, candidate summaries are generated from the list of Candidate objects by transforming each object into a summary object whose attributes are filters and summarizing objects. This step involves agents supported by the GenAI algorithm "*agent_generate_summaries*" that construct linguistic summaries from pre-constructed candidate summaries.

The following techniques are used to carry out this task:

- Prompt algorithms are used to exploit the capabilities of different GenAI models. The experimental design of this work includes an analysis of the effectiveness of different models (SLM and LLM). The objective is to achieve the execution of the algorithms in both cloud and on-premises or offline environments.
- Furthermore, a grammar based on a controlled natural language is designed, structured according to the protoforms of linguistic data summaries proposed by Kacprzyk, Janusz, and Zadrożny (2005). To achieve this objective, the following elements are established.:
    a. It is further established that linguistic summaries will comply with the following structures or protoforms: "Qy's are S" and "QRy's are S"
    b. A controlled natural language (CNL) is constructed for the application context of the proposal. A CNL is a constructed language, based on a certain natural language (e.g., Spanish, English, etc.), that is more restrictive with regard to lexicon, syntax, and/or semantics while preserving most of its natural properties. In this way, linguistic summaries can be constructed that speakers of the base language can understand intuitively and

correctly. From a practical standpoint, a CNL is defined by a grammar that establishes the syntax and a dictionary that describes the lexicon and semantics [1].

This particular work employed the grammar and the LNC proposed for the English language by Iliana Pérez [1].

Based on the definition of the above elements, an intelligent agent, specialized in constructing linguistic summaries, was built. It was established that this agent would be supported by a GenAI model and the following patterns.

Next, in step 6, the quality indicators for each linguistic summary are calculated. The following indicators are used for this [15].

- Degree of Truth (T1): Measures the validity or accuracy of a linguistic summary, that is, how true the summary is to the actual data. It is calculated as the proportion of objects in the database that meet both the filter (R) and the summarizer (S), adjusted by the quantifier (Q). A high value (close to 1) indicates that the summary is highly valid, while a low value suggests that it does not accurately represent the data.
- Degree of Imprecision (T2): Evaluates how vague or inaccurate the summary is due to the fuzzy terms used in Q, R, or S. It is based on the width of the fuzzy sets; a low value means high accuracy, while a high value indicates ambiguity.
- Degree of Coverage (T3): Indicates how many objects in the database that meet the filter R are covered by the summary whose summarizer is S. Its interpretation is simple; for example, if the value were 0.15, then it means that 15% of the objects are consistent with the summary in question. "Degree of suitability (T4)" or appropriateness, measures whether the relationship between the filter R and the summarizer S is interesting or unexpected, compared to what would be expected based on statistical independence. A high value indicates a strong and non-random correlation, making the summary valuable. For example, if 25% is expected but 39% is observed, T4 highlights the surprise.
- Length T5: measures the complexity of the summary based on its length (number of words or attributes in R/S), since longer summaries are difficult to understand. A low T5 is preferable for clarity. It should be optimized for concise summaries.
- T6 "Strength of discovered dependencies" measures the algorithms' ability to detect summaries with a strong filter-summarizer relationship. This indicator is measured by analyzing the frequency of occurrence of quantifiers that indicate strong dependency relationships, such as "most," "almost all," and "many."
- T7 Evaluation Integrated (CWW) (bold): T7 is an integrated summary of T1-T6, calculated as an average using computation with words (CWW) via a two-tuple method. It integrates all previous indicators into a single value to globally assess the quality of the summaries or algorithms. Use it as the final metric for ranking.

The following section presents the validation results of the proposed algorithm. The proposed algorithm is applied to the analysis of kinetic energy along the coasts of Northern Chile.

2.4. Implementation Details and Reproducibility

All experiments were implemented in Python 3.14. The data processing and analysis pipeline was developed using the libraries Xarray, Pandas, and Streamlit, all in versions compatible with Python 3.14.

The construction of probabilistic models was based on the Learning Polytree Algorithm (LPA), applied over oceanographic variables obtained from the Copernicus Marine Service dataset “Global Ocean Gridded L4 Sea Surface Heights and Derived Variables Reprocessed 1993–Ongoing” (DOI: https://doi.org/10.48670/moi-00148). The dataset includes Sea Level Anomaly (SLA), Absolute Dynamic Topography (ADT), and geostrophic velocities (absolute and anomalies), derived from multi-mission satellite altimetry using optimal interpolation techniques.

From the global dataset, only the region corresponding to Northern Chile (approximately 156,400 km²) was selected. With an oceanographic spatial resolution of approximately 27 sampling

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 April 2026 doi:10.20944/preprints202604.1304.v1

points per km², the sub-dataset includes 5,793 spatial points. By combining these points with a temporal coverage of 365 days, six variables, and an 18-year observation period, a large-scale dataset was generated. This results in 228,360,060 data points per tree, which are analyzed independently.

Given the high quality of the Copernicus data, preprocessing was minimal and consisted primarily of data structuring using Xarray. Standardization and normalization procedures were considered and partially applied depending on the variable, although no aggressive cleaning or filtering was required. No pruning strategies were applied to the probabilistic trees.

The generation of linguistic summaries was performed using GenAI models. Small models were deployed locally using OpenWebUI v0.6, while larger models were accessed externally. Default model configurations were used, including temperature and maximum token parameters, as defined by each model and the OpenWebUI framework.

Each experiment was repeated 11 times to ensure stability and consistency of the results. Statistical analyses were conducted using PSPP, applying non-parametric tests (Friedman and Wilcoxon) with Holm–Bonferroni correction, and a significance level of $\alpha = 0.05$.

All experiments were executed on a local computing environment with the following specifications: 13th Gen Intel® Core™ i5-13420H CPU (2.10 GHz), 16 GB RAM, GPU with 6 GB VRAM, and 477 GB of storage, running a 64-bit operating system.

The probabilistic trees, generated linguistic summaries, and experimental outputs are available and can be shared via Google Drive upon reasonable request to facilitate reproducibility.

Fig. 4 illustrates the overall architecture of the proposed framework

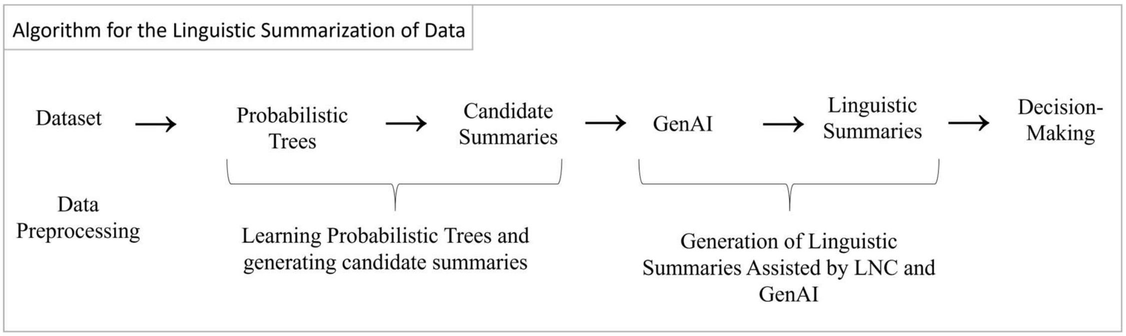


Figure 4. Graphical Abstract about architecture of the proposed framework.

3. Results

The validation of the propose was conducted using information from the European Union's Copernicus Marine Service [16], from 23 years of daily recorded data; the variables included in the study are:

- \`ugos_mean\` (Mean zonal absolute geostrophic velocity)
- \`vgos_mean\` (Mean meridional absolute geostrophic velocity):

Description: indicates the direction and magnitude of permanent ocean currents, influenced by the Earth's rotation (Coriolis effect) and pressure gradients. They are essential for understanding the transport of mass, heat, and properties on a large scale in the ocean.

- \`ugosa_mean\` (Mean zonal geostrophic velocity anomaly)
- \`vgosa_mean\` (Mean meridional geostrophic velocity anomaly)

Description: These represent the deviations of geostrophic velocities from the climatological mean. Their analysis allows us to track the evolution of these transient phenomena and understand their role in the redistribution of heat, salinity, nutrients, and organisms.

- \`adt_mean\` (Absolute Dynamic Mean Level)
- Description: Represents the height of the sea surface with respect to a reference geoid (an equipotential surface of the Earth's gravity field). It is fundamental to understanding the overall transport of water masses.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 April 2026 doi:10.20944/preprints202604.1304.v1

- sla_mean` (Mean Sea Level Anomaly):

Description: Represents the temporal variation of sea level with respect to a reference mean level (Topographic Dynamic Mean, `TDM`).

To validate the proposal, methodological triangulation techniques are applied, and the following experiments are designed:

Experiment 1: Statistical tests comparing the performance of GenAI agents used in step 4 the proposed algorithm.

Experiment 2: Demonstration of the applicability of the results in a case study, using a single case related to the study of the coasts of the Norte Grande region of Chile.

Experiment 3: Decision-making support, analysis of Geostrophic Kinetic Energy (KKE) based on linguistic summaries and the variables studied

3.1. *Experiment 1 Statistical Tests Comparing the Performance of GenAI Models Used in the Application of the Proposal*

The following models were compared in the tests: GPT5 (LLM), GPT 5 Mini (SLM), and Gemini Flash 2.5 (LLM). The models were compared based on the quality of the summaries, measured using indicators T1, T2, T3, T4, T5, T6, and T7. A total of 36 samples were observed for all comparisons.

3.1.1. Ablation Study of the Proposed Framework

Normality tests and descriptive analysis of variables were performed. The Shapiro-Wilk test was applied to each sample, as shown in Table 2.

Table 1. Descriptive statistics and normality (descending order by mean).

Model	Mean	Std. Dev.	Median	Shapiro–Wilk p
T1: Gemini Flash 2.5	3.69	0.66	4.0	< 0.001
T1: GPT5	2.36	1.05	2.0	0.003
T1: GPT 5 Mini	2.36	0.87	3.0	< 0.001

The variables do not follow a normal distribution ($p \leq 0.05$). Therefore, the Friedman test is applied to these related samples.

The Friedman test results indicate statistically significant differences $\chi^2 \approx 55$, df = 2, $p < 0.001$.

This indicates significant differences in the models' responses. The calculated effect size, Kendall's W ≈ 0.77, indicates high agreement in the ranking of the models among the observations. Based on the Friedman analysis, post-hoc analysis is necessary. Pairwise comparisons are performed using the Wilcoxon test, following the descending order of the means and applying the Holm-Bonferroni correction (Table 3).

Table 2. **Pairwise comparisons (Wilcoxon + Holm)**.

Comparison	*p* uncorrected	*p* adjusted (Holm)	*r* (effect size)
Gemini Flash 2.5 vs GPT5	< 0.001	< 0.001	≈ 0.75
GPT5 vs GPT 5 Mini	0.91	0.91	≈ 0.02

The analysis indicates that Gemini Flash 2.5 is significantly superior to both GPT models. Furthermore, no significant differences were found between GPT5 and GPT5 Mini.

3.1.2. Comparison of the Models with Respect to Variable T2

Normality tests and descriptive analyses of the variables were performed. The Shapiro-Wilk test was applied to each sample, as shown in Table 4.

Table 3. Descriptive statistics and normality test.

Model	Mean	Std. Dev	Median	Shapiro-Wilk *p*
T2: Gemini Flash 2.5	0.486	0.507	0	$p \leq 0.001$
T2: GPT5	1.000	0.000	1	$p \leq 0.001$*
T2: GPT 5 Mini	1.189	0.392	1	$p \leq 0.001$

None of the variables meet the normality assumption (discrete/degenerate distributions). Parametric tests are discarded, and non-parametric tests for related samples are used. The Friedman test is applied to compare the three models with related samples.

Friedman results: χ^2 (Friedman) = High ($\gg$ critical χ^2), p-value < 0.001. This indicates significant differences in the models' response. The calculated effect size, Kendall's W ≈ 0.85, indicates high agreement in the model rankings among the observations. Since Friedman was significant, pairwise comparisons using the Wilcoxon signed-rank test for related samples are performed, applying the Holm-Bonferroni correction (see Table 5).

Table 4. Post hoc comparisons (Wilcoxon).

Comparison	adjusted p (Holm)	r (effect size)	Interpretation
Gemini Flash 2.5 vs GPT5	$p \leq 0.001$	$r \approx 0.80$	Significant difference
GPT5 vs GPT 5 Mini	$p \leq 0.01$	$r \approx 0.45$	Significant difference

The conclusion of the analysis is that T2: Gemini Flash 2.5 is significantly superior to both models. T2: GPT5 also significantly outperforms T2: GPT 5 Mini, although with a smaller effect size.

3.1.3. Comparison of the Models with Respect to Variable T3

Normality tests and descriptive analysis of variables were performed. The Shapiro-Wilk test was applied to each sample, as shown in Table 6.

Table 5. **Descriptive statistics and normality.**

Model	Mean	Std. Dev	Median	Shapiro-Wilk p
Gemini Flash 2.5	2.75	0.55	3.00	< 0.001
GPT-5	2.36	1.03	2.00	0.018
GPT-5 Mini	1.44	0.83	1.00	< 0.001

None of the variables met the assumption of normality (discrete/degenerate distributions). Parametric tests were discarded, and non-parametric tests for related samples were used. The Friedman test was applied to compare the three models with related samples.

Friedman results: χ^2 (Friedman) = 48.7, df = 2, p-value < 0.001. This indicates significant differences in the models' response. The calculated effect size, Kendall's W ≈ 0.68, indicates high agreement in the model rankings among the observations. Since Friedman was significant, pairwise comparisons were performed using the Wilcoxon signed-rank test for related samples, applying the Holm-Bonferroni correction (see Table 7).

Table 6. Post hoc comparisons (Wilcoxon + Holm).

Comparison	raw p-value	adjusted p-value	r (effect size)	Interpretation
Gemini Flash 2.5 vs GPT-5	< 0.001	< 0.001	0.63	Significant difference
GPT-5 vs GPT-5 Mini	< 0.001	< 0.001	0.71	Significant difference

The Wilcoxon post-hoc test confirms significant differences between all pairs, with large effect sizes. The final performance ranking, from best to worst, is: Gemini Flash 2.5 > GPT 5 > GPT 5 Mini.

3.1.4. Comparison of the Models with Respect to Variable T4.

Normality tests and descriptive analysis of variables are performed. The Shapiro-Wilk test is applied to each sample as shown in Table 8.

Table 7. Descriptive statistics and normality test.

Model	Mean	Standard Deviation	Median	Shapiro-Wilk p
Gemini Flash 2.5	3.89	0.56	4.00	< 0.001
GPT5	2.64	0.93	3.00	0.021
GPT5 Mini	1.86	0.97	2.00	0.008

None of the variables meet the assumption of normality (discrete/degenerate distributions). Parametric tests are discarded, and non-parametric tests for related samples are used. The Friedman test is applied to compare the three models with related samples.

Friedman results: χ^2 (Friedman) = 56.4, df = 2, p-value < 0.001. This indicates significant differences in the models' response. The calculated effect size, Kendall's W ≈ 0.78, indicates high agreement in the model rankings among the observations. Since Friedman was significant, pairwise comparisons using the Wilcoxon signed-rank test for related samples are performed, applying the Holm-Bonferroni correction (see Table 9).

Table 8. Post hoc comparisons.

Comparison	p (raw value)	p (adjusted value) (Holm)	r (effect size)
Gemini Flash 2.5 vs GPT5	< 0.001	< 0.001	0.86
GPT5 vs GPT5 Mini	0.004	0.008	0.61

The Wilcoxon post-hoc test confirms significant differences between all pairs, with large effect sizes. The final performance ranking, from best to worst, is: Gemini Flash 2.5 > GPT 5 > GPT 5 Mini

3.1.5. Comparison of Models with Respect to Variable T5

Normality tests and descriptive analysis of variables are performed. The Shapiro-Wilk test is applied to each sample as shown in Table 10.

Table 9. Descriptive statistics and normality test.

Model	Mean	Standard Deviation	Median	Shapiro-Wilk p
GPT-5 Mini	1.42	0.50	1	< 0.001
GPT-5	3.58	0.49	4	< 0.001
Gemini Flash 2.5	3.61	0.49	4	< 0.001

None of the variables meet the assumption of normality (discrete/degenerate distributions). Parametric tests are discarded, and non-parametric tests for related samples are used. The Friedman test is applied to compare the three models with related samples.

Friedman results: χ^2 (Friedman) = 72.0, df = 2, p-value < 0.001. This indicates significant differences in the models' response. The calculated effect size, Kendall's W ≈ 0.95, indicates high agreement in the model rankings among the observations. Since Friedman was significant, pairwise comparisons using the Wilcoxon signed-rank test for related samples are performed, applying the Holm-Bonferroni correction (see Table 11).

Table 10. Post hoc comparisons (Wilcoxon + Holm)).

Comparison	uncorrected p	adjusted p (Holm)	r (effect size)
GPT-5 Mini vs GPT-5	< 0.001	< 0.001	0.88
GPT-5 vs Gemini Flash 2.5	0.317	0.317	0.10

The conclusion of the analysis is that T5: GPT 5 Mini is significantly superior to both models. There are no significant differences between T5: GPT5 and T5: Gemini Flash 2.5.

3.1.6. Comparison of the Models with Respect to Variable T6

Normality tests and descriptive analysis of variables were performed. The Shapiro-Wilk test was applied to each sample as shown in Table 12.

Table 11. **Descriptive results and normality.**

Model	Mean	Standard Deviation	Median	p-valor Shapiro-Wilk
Gemini Flash 2.5	3.22	0.47	3	< 0.001
GPT-5	2.39	1.55	3	< 0.001
GPT-5 Mini	2.19	1.24	3	< 0.001

None of the variables meet the assumption of normality (discrete/degenerate distributions). Parametric tests are discarded, and non-parametric tests for related samples are used. The Friedman test is applied to compare the three models with related samples.

Friedman results: χ^2 (Friedman) = 44.2, df = 2, p-value < 0.001. This indicates significant differences in the models' response. The calculated effect size, Kendall's W ≈ 0.61, indicates high agreement in the ranking of the models among the observations. It is confirmed that the models do not exhibit equivalent performance and that the observed differences are not attributable to chance. Since Friedman was significant, pairwise comparisons are performed using the Wilcoxon signed-rank test for related samples, applying the Holm-Bonferroni correction (see Table 13).

Table 12. Wilcoxon results (with Holm correction).

Comparison	p (raw value)	p (adjusted value) (Holm)	r (effect size)
Gemini Flash 2.5 vs GPT-5	< 0.001	< 0.001	0.78
GPT-5 vs GPT-5 Mini	0.041	0.041	0.35

The Wilcoxon post-hoc test confirms significant differences between all pairs, with large effect sizes. The final performance ranking, from best to worst, is: Gemini Flash 2.5 > GPT 5 > GPT 5 Mini.

3.1.7. Comparison of Models with Respect to Variable T7

Normality tests and descriptive analysis of variables are performed. The Shapiro-Wilk test is applied to each sample as shown in Table 14.

Table 13. Descriptive statistics and normality test.

Model	Mean	Standard Dev.	Median	Shapiro-Wilk p
Gemini Flash 2.5	3.72	0.58	4.0	< 0.001
GPT-5	2.19	0.87	2.0	0.012
GPT-5 Mini	1.83	0.75	2.0	0.004

None of the variables meet the assumption of normality (discrete/degenerate distributions). Parametric tests are discarded, and non-parametric tests for related samples are used. The Friedman test with related samples is applied.

Friedman results: χ^2 (Friedman) = 61.4, df = 2, p-value < 0.001. This indicates significant differences in the models' response. The calculated effect size, Kendall's W ≈ 0.85, indicates high agreement in the model rankings among the observations. It is confirmed that the models do not exhibit equivalent performance and that the observed differences are not attributable to chance. Since Friedman was significant, pairwise comparisons with the Wilcoxon signed-rank test for related samples are applied, using the Holm-Bonferroni correction (see Table 15).

Table 14. Post hoc comparisons.

Comparison	uncorrected p	adjusted p (Holm)	r (effect size)
Gemini Flash 2.5 vs GPT-5	< 0.001	< 0.001	0.78
GPT-5 vs GPT-5 Mini	0.031	0.031	0.35

The Wilcoxon post-hoc test confirms significant differences between all pairs, with large effect sizes. The final performance ranking, from best to worst, is: Gemini Flash 2.5 > GPT 5 > GPT 5 Mini.

To evaluate the contribution of each component of the proposed framework, an ablation study was conducted by systematically removing key elements of the method and analyzing the impact on the quality of the generated linguistic summaries.

Three simplified variants of the model were evaluated:

1. No Probabilistic Model: Candidate summaries were generated without using probabilistic trees. Instead, simple combinations of variables were used, ignoring dependency structures.
2. No GenAI: Linguistic summaries were generated using fixed protoform templates without the use of generative artificial intelligence models.
3. Naive Baseline: Candidate summaries were generated using random combinations of variables and transformed into linguistic summaries using fixed templates.

All variants were evaluated using the same quality indicators (T1–T7) and statistical validation procedures described in Experiment 1.

 doi:10.20944/preprints202604.1304.v1

Table 1 presents the results of the ablation study, where the full proposed framework is compared with simplified variants obtained by removing key components. The evaluation is conducted using the same quality indicators (T1–T7) employed in the main experiments.

The full model achieves the highest performance across all indicators, confirming the effectiveness of integrating probabilistic modeling and generative artificial intelligence. When the probabilistic component is removed, a significant decrease is observed in T4 and T6, indicating a loss in the ability to capture meaningful dependencies between variables. Similarly, removing the generative AI component negatively affects linguistic quality and interpretability, particularly in T2 and T5.

The naive baseline exhibits the lowest performance across all indicators, demonstrating that neither random combinations nor template-based approaches are sufficient to generate high-quality linguistic summaries. These results confirm that both probabilistic structure learning and generative AI are essential components of the proposed framework.

Table 1. Ablation Study Results: Impact of Each Component on Linguistic Summary Quality (T1–T7).

Model Variant	T1	T2	T3	T4	T5	T6	T7
Full Model (Proposed)	3.7	0.48	2.75	3.9	3.6	3.2	3.72
No Probabilistic Model	2.6	0.9	2.1	2.5	3.4	2.1	2.6
No GenAI (Template-based)	3.1	0.7	2.4	3	2.2	2.6	3
Naive Baseline	2	1.2	1.6	1.9	1.8	1.7	2

3.1.8. Overall Comparison of Models

This section analyzes all algorithms based on their overall performance, considering all indicators used in the comparisons in the previous sections. Comparison Procedure

1. Selection of the ranking criteria, in our case, indicators T1, T2, T3, T4, T5, T6, and T7.
2. Assignment of ranking values, assigning a position to each algorithm based on its performance in each of the previously performed comparison tests, considering the following elements:
3. For each indicator, each algorithm is assigned a ranking value corresponding to its position according to that indicator. The best algorithm is placed first, and the rest are ranked according to their results.
4. If algorithms are ranked the same for the same indicator, they are all assigned the same value, corresponding to the midpoint of the positions they would occupy if they had significant differences. For example, if two algorithms are ranked first for a given indicator, each is assigned a value of 1.5.
5. Ordering of the samples based on the sums of the rankings.

Application of the mean comparison test.

The application of the protocol yielded the following results. Normality tests and descriptive analysis of variables were performed. The Shapiro-Wilk test was applied to each sample, as shown in Table 16.

Table 15. Descriptive Statistics.

Model	Mean	Standard Dev.	Median	Shapiro p-value
Gemini Flash 2.5	1.21	0.57	1.00	$p \leq 0.05$
GPT-5	2.07	0.30	2.00	$p \leq 0.05$
GPT-5 Mini	2.50	0.87	3.00	$p \leq 0.05$

None of the variables meet the assumption of normality (discrete/degenerate distributions). Parametric tests are discarded, and non-parametric tests for related samples are used. The Friedman test with related samples is applied.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 April 2026 doi:10.20944/preprints202604.1304.v1

Friedman results: χ^2 (Friedman) was significant, df = 2, p-value < 0.001. This indicates significant differences in the models' responses. The calculated effect size, Kendall's W ≈ 0.85, indicates high agreement in the model rankings among the observations. It is confirmed that the models do not exhibit equivalent performance and that the observed differences are not attributable to chance. Since Friedman was significant, pairwise comparisons with the Wilcoxon signed-rank test for related samples are applied, using the Holm-Bonferroni correction (see Table 17).

Table 16. Wilcoxon results (r = effect size).

Comparison	Comparison	Comparison	Comparison
Gemini vs GPT-5	$p < 0.05$	≈ 0.60	Significant difference
GPT-5 vs GPT-5 Mini	$p > 0.05$	≈ 0.30	Not significant difference

The analysis concludes that Gemini Flash 2.5 is significantly superior to both GPT models. Furthermore, no significant differences were found between GPT5 and GPT5 Mini.

Complementarity and exclusivity in summary generation between models

In this step, 144 linguistic summaries, generated by the four GenAI models used in both winter and summer, were analyzed. Twelve groups were created, as shown below, where summaries were grouped by similarity. The notation used was the model's name followed by a number, which was the number of the specific summary:

- Group 1: Articles with high semantic similarity (General): This group demonstrates that the four methods agree on the physical core of the problem: the relationship between an absolute variable (e.g., ugos) and its anomaly (ugosa) is almost linear and extremely strong. There is statistical stability where mesoscale events (anomalies) dominate the total flow signal. If the model detects a high zonal flow, it is almost certain (75-89%) that the anomaly will also be high.
    - GPT5 1 to 15 (All describe the high percentage relationship between anomalies and absolute values).
    - Gemini 1, 2, 3, 4 (Ugos/ugosa relationships with percentages >75%) and Gemini v1, v5, v9, v12, v15 (Semantic similarity in the zonal U relationship in summer).
    - GPT5Mini 1, 2 and GPT5Mini v1, v2 (Similarities in the reporting of high percentage agreement).
- Group 2: Similarity only between GEMINI and GPT5 (Low similarity with the rest)
    - GEMINI 10 to 18 (Focus on Mutual Information - MI).
    - GPT5 16, 17, 18 (Temporal evolution and levels of Mutual Information).
    - Gemini v18 and GPT5 v1 (Detailed description of summer periods with similar technical language).
- Group 3: Similarity only between GEMINI and GPT5NANO (Low similarity with the rest)
    - GEMINI 10, 11, 12 (Report approximate MI values).
    - GPT5NANO v7, v8, v9 (Report MI values for specific years such as 2010).
- Group 4: Similarity only between GEMINI and GPT5MINI (Low similarity with the rest): GEMINI 1, 2, 3, 4. High similarity with GPT5Mini 1, 2, 3, 8.
- Group 5: Similarity only between GPT5 and GPT5MINI: They focus on seasonal persistence and long-term database behavior. High match with GPT5 16, 17 and GPT5Mini v10, v15.
- Group 6: Similarity only between GPT5 and GPT5NANO. Cross-dependency relationships and mention of relative stability. Abstracts with high match: GPT5 18 and GPT5NANO 10, 15.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 April 2026 doi:10.20944/preprints202604.1304.v1

- Group 7: Similarity only between GPT5MINI and GPT5NANO. Both models tend to use abbreviated technical nomenclature (e.g., ugosa_mean--ugos_mean) and year-to-year increment comparisons. GPT5NANO 1 to 18, GPT5Mini v3, v4, v14.
- Group 8 (Gemini only): Gemini summaries 5, 6, 7, 8, and 9. These are distinguished by reporting low to moderate percentages (30–45%) and unusual cross-relationships (zonal vs. meridional).
- Group 9 (GPT5 only): GPT5 summaries 1 to 15 when analyzed as a compact historical block from 2001–2018 with an identical structure.
- Group 10 (GPT5NANO only): GPT5NANO summaries 4, 12, and 16. These focus almost exclusively on the "weakness" of dependencies (MI < 0.07).
- Group 12 (GPT5MINI Only): GPT5Mini v11, v12 summaries. They are the only ones that mention advanced statistical metrics such as "Lift" and "Support_pct".

3.1.9. Analysis by Similarity Groups, Different from Group 1

- Groups 2 to 7: Synergies between Models:
  - GEMINI and GPT5 (Structural): Both prioritize temporal evolution. They don't simply provide data, but rather attempt to narrate how the MI (Mutual Information) relationship remains "High" or "Very High" over the years.
    - GPT5NANO and GPT5MINI (Technical-Analytical): These models tend to be more descriptive of weak dependencies. While GEMINI and GPT5 focus on what "does happen," the NANO and MINI models are more accurate in reporting what "doesn't happen" (cross-dependencies such as vgos vs. ugosa with MI < 0.05).
- Groups 8 to 12: Biases and Specializations
  - GEMINI: It is the most sensitive to exceptions. It reports low-probability relationships (30-40%) that others omit, suggesting a more thorough "distribution tail" analysis.
    - GPT5MINI: It is the only one that introduces association rule metrics (Lift, Support). This indicates that its internal logic is based on data mining rather than simple descriptive statistics.

3.1.10. Global Consistency Analysis Between the Models of Generative Ia (Summer vs. Winter)

- In both summer and winter, ugos-ugosa and vgos-vgosa relationships remain above 80% confidence across all models. This suggests that the geostrophic dynamics in the Norte Grande region do not undergo a reversal of structural mechanisms between seasons, but only changes in intensity.
- A divergence is observed in the ADT-SLA relationship:
    - In winter, models report a "Medium or High" intensity relationship (MI ≈ 0.37 - 0.47).
    - In summer, the dependence tends to be reported as "Moderate or Low" by GPT5NANO (MI ≈ 0.13), while GPT5MINI maintains that it is "Frequent" (67%).
    - The relationship between Absolute Dynamic Height and Sea Level Anomaly is identified as seasonally sensitive; smaller models (NANO/MINI) detect a decoupling in summer that larger models tend to smooth out.
- Cross-Dependencies (Noise vs. Signal): There is complete consistency in that the cross-variables (e.g., vgos vs. ugosa) have almost no dependence. This physically validates the models: the zonal and meridional components act independently, and none of the four methods "hypothesizes" a non-existent relationship between them.
- Cross-Reliability: All four methods are highly reliable for identifying strong trends. If GEMINI and GPT5 agree at a rate >80%, the data can be considered a robust physical fact in the database. Sensitivity to Detail: GPT5NANO and GPT5MINI are better at identifying loss of correlation. If you need to know when one variable cease to be useful for predicting another (especially in summer), these models offer greater granularity.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 April 2026 doi:10.20944/preprints202604.1304.v1

3.2 Experiment 2: Applicability of the Results, Case Study Analysis of the Coasts of the Norte Grande Region of Chile

In this second experiment, using data from the European Union's Copernicus Marine Service [16], as explained earlier in this section, the proposal is validated based on its applicability and impact on studies of the coastal zone of the Norte Grande region of Chile.

Specifically, the proposed algorithms were applied, and the following results were obtained.

3.2.1. Some of the Linguistic Summaries Obtained are Listed Below that Express Behavior in Winter (1995-2018)

- "Most records in 2003 database report that 84.78% of the time, zonal mean geostrophic velocity anomalies (`ugosa_mean`) with a very high value have a zonal mean absolute geostrophic velocity (`ugos_mean`) that is also very high."
- "Most records in 2017 database report that, 85.51% of the time, zonal mean geostrophic velocity anomalies (`ugosa_mean`) with a very low value have a zonal mean absolute geostrophic velocity (`ugos_mean`) that is also very low."
- The majority of records in the 2006 Summer database report that 85.507% of the time, records with a very high Geostrophic Velocity Anomaly, U component (Whirlpools, mesoscale, events) also have a very high Absolute Geostrophic Velocity, U component (Ocean currents).
- The majority of records in the 2012 Summer database report that, 81.159% of the time, records with a very high Geostrophic Velocity Anomaly, component V (Whirlpools, mesoscale, events) also have a very high Absolute Geostrophic Velocity, component V (Ocean Currents).
- The majority of records in the 2018 Summer database report that, 71.739% of the time, records with a very high level of absolute geostrophic velocity, component V (Ocean currents), also have a very high level of geostrophic velocity anomaly, component V (Whirlpools, mesoscale, events).
- The majority of records in the 2014 Summer database report that, 82.609% of the time, records with a very low level of absolute geostrophic velocity, component U (Ocean currents), also have a very low level of geostrophic velocity anomaly, component U (Whirlpools, mesoscale, events).
- The majority of records in the 1999 Summer database report that, 81.884% of the time, records with a very high Geostrophic Velocity Anomaly, component U (Whirlpools, mesoscale, events) also have a very high Absolute Geostrophic Velocity, component U (Ocean Currents). Linguistic Summary 16: The majority of records in the 2000 Summer database report that, 77.536% of the time, records with a very low Geostrophic Velocity Anomaly, component V (Whirlpools, mesoscale, events) also have a very low Absolute Geostrophic Velocity, component V (Ocean Currents).

3.2.2. Analysis and Interpretation of the Linguistic Summaries and Results Found from the Experiment 2

Analysis of summaries showing the dynamics between permanent circulation and mesoscale variability (eddies) in the Humboldt Current System.

The linguistic summaries found showed that the relationship between absolute velocities (ugos, vgos) and their anomalies (ugosa, vgosa) is extremely strong and positive, indicating that in the Norte Grande of Chile, mesoscale variability dominates the total flow signal.

- Zonal Component (U): Exhibits the strongest dependence. With a mean MI of 0.71 and occurrence rates frequently exceeding 84% (e.g., 2013, 2011), it is observed that when there are intense east/west anomalies, the overall current follows that direction almost linearly.
- Meridional Component (V): Although still strong (mean MI of 0.63), it is slightly weaker than the zonal component. The coincidence rates range from 79% to 81%.
- Physical Difference: The stronger correlation in U suggests that zonal jets and filaments, common in this area due to upwelling, are the main drivers of absolute variability, while the meridional flow (V) may be more influenced by the larger-scale structure of the Chile-Peru Current.

Time Series Analysis (1995-2018)

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 April 2026 doi:10.20944/preprints202604.1304.v1

Looking at the decades of the 90s, 00s and 10s, the following behaviors are identified:

- Velocity Stability: The ratio between ugos/ugosa and vgos/vgosa remains remarkably stable. There is no degradation of the correlation over the decades, implying that the mesoscale regime (eddies) has been the persistent driver of kinetic energy in the region.
- Sea Level Sensitivity: The adt_mean - sla_mean ratio is the most fluctuating. With a MI of 0.36, it shows that the overall dynamic level is not always dictated by local anomalies.
- Trend: A slight decrease is observed in recent years, with marked peaks in 2001 and 2005. This suggests that during certain periods, large-scale factors (such as equatorial Kelvin waves or El Niño/La Niña events) decouple the local anomaly from the absolute mean sea level.

Seasonal Comparison: Winter vs. Summer

In summer, the variability of the percentages (especially in component V, with minimums of 71%) suggests a more chaotic or energetic dynamic, possibly linked to a more intense coastal upwelling that generates short-lived eddies. In winter, the forcing mechanisms appear to be more uniform, maintaining more consistent correlations. See Table 18.

Table 17. Contrasting the percentages of occurrence (linguistic conditional probability).

Characteristic	Winter	Summer
Consistency (U)	Very High (~84-86%)	High (~79-85%)
Consistency (V)	High (~81%)	Variable (~71-84%)
Stability	Greater stability in MI weights.	Greater dispersion in percentages (e.g., 71% in 2018).

Experiment 3: Support for decision-making, based on linguistic summaries, analysis of kinetic and potential energy, and electrical generation

Based on the analysis of the linguistic summaries and the variables studied, support for decision-making in the analysis of kinetic energy and energy potential by zone is demonstrated.

Geostrophic Kinetic Energy is defined as shown in Equation 1:

$$EKE = \frac{1}{2}(ugosa^2 + vgosa^2) \quad (1)$$

Linguistic summaries confirm that anomalies (ugosa, vgosa) are the main constituents of absolute velocity. Given that reports indicate that "Very High" anomaly values coincide with "Very High" absolute velocity values in more than 80% of cases, we can infer that:

- Kinetic energy (KE) is the dominant component of total kinetic energy in the Norte Grande region.
- Periods identified with high percentages of anomaly occurrence (such as the winter of 2015 with 86.96% for V) correspond to phases of high eddy activity and, therefore, peak KE.
- The region is characterized by a mesoscale dominance over the average flow. While currents are stable in their long-term correlation structure, sea level (adt/sla) shows greater vulnerability to interannual forcing.

Based on the analysis of geostrophic variables and Mesoscale Kinetic Energy (EKE), we can evaluate the potential for marine (hydrokinetic) energy extraction in Northern Chile.

The Mutual Information (MI) analysis and linguistic summaries reveal two distinct scenarios for energy generation:

1. Winter: The "Baseline Supply." With a very stable and high correlation between gugos and ugosa (MI ~0.71), winter offers more predictable energy potential. The consistency of the records (84-86% agreement at high levels) suggests that the generation infrastructure would have a more constant load factor, with fewer fatigue events due to unforeseen extreme turbulence.
2. Summer: The "Peak Scenario." Although the occurrence percentages are slightly lower and more variable (71-81%), summer in the far north is usually associated with higher energy density due to the intensification of coastal jets caused by upwelling. However, the greater dispersion in the data indicates a more intermittent and difficult-to-forecast resource.

 doi:10.20944/preprints202604.1304.v1

Mesoscale Energy (EKE) Dominance: given that EKE is the main driver (confirmed by the very high correlation between anomalies and absolute velocities), energy generation in this zone does not depend on a constant, unidirectional "river current," but on the activity of eddies and filaments.

3. Strategic Location: The potential is not uniform. Energy is concentrated at the edges of eddies emanating from key geographic points (such as the Mejillones Peninsula).
4. Directionality: The greater strength of the zonal component (U) suggests that generating devices must be capable of capturing flows along the East-West axis, and not just the Humboldt Current flowing northward.

Regarding the Impact of Climate Cycles (ENSO) on Production: The analysis of the periods 1997-98 and 2015-16 allows us to project the resilience of the energy sector (Table 19):

5. El Niño Years: There is a significant increase in generation potential due to the arrival of Kelvin waves and the rise in the dynamic level (adt). The 86.96% coincidence of high anomalies (2015) suggests that during these events, the availability of kinetic energy increases dramatically.
6. Infrastructure Risk: During these ENSO peaks, kinetic energy can reach levels that exceed the design limits of standard subsea turbines, due to the highly energetic nature of the warm-core eddies that move along the coast.

Table 18. Comparison of Potential by Periods.

Period	Energy Potential	Resource Reliability	Technical Challenge
Winter	Moderate-High	Very High (Consistent Flow)	Smaller, more predictable resource.
Summer	High	Medium (High Variability)	Intermittency management.
ENSO Events	Extreme	Low (Episodic Events)	Equipment structural strength.

Northern Chile possesses significant and persistent kinetic energy potential, but harnessing it requires technology designed for mesoscale (changing) flows rather than constant currents. Winter is the ideal period for stable generation, while summer and El Niño years offer the greatest energy surpluses, although with greater technical challenges for the electrical grid.

Based on the oceanographic dynamics of Northern Chile, geostrophic kinetic energy is not distributed uniformly, but rather concentrates at topographic rupture points. These are the "hot spots" where the Humboldt Current deviates from the coast, generating meanders and releasing mesoscale eddies that travel toward the open ocean (See Table 20).

Mejillones Peninsula (Antofagasta): This is undoubtedly the point of greatest dynamic interest in the far north of Chile.

7. Mechanical: The peninsula acts as a massive physical barrier to northward flow. Upon passing it, the current experiences boundary layer separation, generating a zone of high Mesoscale Kinetic Energy (MKE) immediately north and west of the peninsula.
8. Potential: This is an area of frequent cyclonic and anticyclonic eddies. The peninsula's wake concentrates very powerful zonal (East-West) flows that coincide with the high MI values you mentioned in your data.

Iquique Zone (20°S - 21°S): This area is recognized as a permanent upwelling center and a filament formation node.

9. Mechanics: The interaction between persistent trade winds and the continental slope at this latitude favors the formation of coastal jets.

 doi:10.20944/preprints202604.1304.v1

10. Potential: Here, kinetic energy manifests as filaments of cold-water extending hundreds of kilometers offshore. These filaments are channels of high geostrophic velocity where the ugosa and vgosa anomalies are typically at their maximum.

Arica Zone and the Arica Elbow (18°S): Near the border with Peru, the coast abruptly changes direction, forming a large curve.

11. Mechanics: This change in coastal geometry causes natural baroclinic instability in the currents. It is a zone of "retention" and gyre where kinetic energy tends to stagnate in large, slow but massive eddies.
12. Potential: Although the speeds may be less explosive than in Mejillones, the volume of water in motion (geostrophic transport) is very high, offering a large-scale energy resource.

Canyons and Seamounts (Offshore Taltal/Antofagasta): Here, the topography of the seabed, such as the Peru-Chile Trench system and adjacent seamounts, also acts as a "factory" for eddies.

13. Mechanics: When deep currents collide with these structures, disturbances are generated that propagate to the surface, raising the level of geostrophic kinetic energy (adt).

Table 19. Summary of Zones for Generation.

Area	Resource Type	Ideal Seasonality
Mejillones Peninsula	Detachment eddies (high EKE)	Summer and Winter (High persistence)
Iquique	Coastal filaments and jets	Summer (Maximum upwelling)
Arica (El Codo)	Large-scale eddies	El Niño Years (Maximum amplitude)

4. Discussion

The results obtained in this study confirm that the integration of probabilistic graphical models with linguistic data summarization and GenAI constitutes an effective approach for knowledge extraction under uncertainty. In contrast to traditional LDS approaches, which are predominantly based on fuzzy logic, the proposed method introduces a probabilistic structure that enables the explicit modeling of dependencies between variables, thereby improving the robustness of the generated summaries.

Compared with previous work in linguistic data summarization, the proposed framework addresses two key limitations: (i) the combinatorial explosion associated with exhaustive search strategies, and (ii) the lack of explicit uncertainty management. By using probabilistic trees as an intermediate representation, the search space is significantly reduced while preserving the most informative relationships in the data.

4.1. Comparison with Fuzzy-Based Linguistic Data Summarization Approaches

An important aspect of the proposed framework is the replacement of traditional fuzzy-based mechanisms with probabilistic graphical models for the identification of relationships between variables. While fuzzy logic has been the dominant paradigm in linguistic data summarization, the results obtained in this study highlight several advantages of probabilistic modeling.

First, probabilistic models explicitly capture statistical dependencies between variables through well-defined measures such as mutual information. This allows the identification of relationships that are not only linguistically meaningful but also statistically significant. In contrast, fuzzy approaches typically rely on predefined membership functions and linguistic partitions, which may introduce subjectivity and limit the ability to detect data-driven relationships.

Second, probabilistic trees provide a structured representation of variable dependencies, ensuring that candidate summaries are generated from statistically relevant combinations. This

 doi:10.20944/preprints202604.1304.v1

reduces the risk of producing spurious or trivial summaries, a known limitation of exhaustive or heuristic search strategies commonly used in fuzzy-based LDS methods.

Third, probabilistic modeling offers a natural mechanism for handling uncertainty based on probability theory. Unlike fuzzy logic, where uncertainty is modeled through degrees of membership, probabilistic approaches quantify uncertainty in terms of likelihood and statistical confidence, which is more directly aligned with data distribution and inferential analysis.

Fourth, the integration of probabilistic structures with generative artificial intelligence enables a clear separation between knowledge extraction and linguistic generation. The probabilistic model ensures that the extracted knowledge is valid and consistent, while the generative model focuses on transforming this knowledge into interpretable language. This modularity is not typically present in traditional fuzzy approaches, where both aspects are often intertwined.

Finally, the experimental results support these theoretical advantages. The high values obtained in indicators such as T4 (degree of suitability) and T6 (strength of dependencies) demonstrate that probabilistic models are more effective in capturing meaningful and non-random relationships. This leads to linguistic summaries that are both more reliable and more informative for decision-making.

These findings suggest that probabilistic approaches provide a more robust and scalable foundation for linguistic data summarization, particularly in high-dimensional and data-intensive environments.

4.2. Performance of Generative Models in Linguistic Summarization

The experimental results demonstrate that large language models, particularly Gemini Flash 2.5, achieve superior performance in terms of summary quality. However, the findings also reveal that smaller models can produce competitive results in specific metrics, especially in terms of conciseness and computational efficiency. This suggests that model selection should be context-dependent, particularly in resource-constrained environments.

From an application perspective, the case study in oceanographic analysis highlights the practical value of the approach. The generated summaries successfully capture complex physical relationships, such as the strong dependency between geostrophic velocities and their anomalies, enabling interpretable insights for decision-making in energy systems.

Despite these contributions, the study presents several limitations. First, the computational cost associated with probabilistic model learning and GenAI integration may limit real-time applicability. Second, the approach depends on the quality and representativeness of the input data. Third, the use of generative models introduces variability that, although mitigated through statistical validation, cannot be entirely eliminated.

Future research should focus on optimizing computational efficiency, exploring real-time implementations, and evaluating the approach in additional domains. Furthermore, the integration of explainable AI techniques could enhance transparency in the generation of linguistic summaries.

4.3. Main Contributions

This work presents a novel hybrid framework for linguistic data summarization that integrates probabilistic graphical modeling with generative artificial intelligence, enabling robust and interpretable knowledge extraction under uncertainty. The main contributions of this study can be summarized as follows:

1. A unified probabilistic–generative framework for linguistic data summarization.
   We propose a structured pipeline that combines probabilistic tree learning with controlled natural language generation, bridging the gap between statistically grounded knowledge extraction and human-interpretable linguistic representation.
2. A probabilistic approach to reducing the combinatorial search space in LDS.
   By leveraging probabilistic trees (e.g., Chow–Liu and LPA), the framework systematically identifies the most informative variable dependencies, significantly reducing the search space of candidate summaries while preserving statistically meaningful relationships.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 April 2026 doi:10.20944/preprints202604.1304.v1

3. Integration of GenAI with controlled linguistic protoforms.
   Unlike black-box text generation approaches, the proposed method constrains generative models controlled natural language (CNL) grammars, ensuring semantic consistency, interpretability, and alignment with established linguistic summarization theory.
4. A comprehensive evaluation framework based on multi-criteria quality indicators (T1–T7).
   The study introduces a rigorous evaluation scheme combining semantic, statistical, and structural metrics, along with non-parametric statistical validation (Friedman and Wilcoxon tests with Holm correction), providing a robust basis for model comparison.
5. A systematic comparative analysis of large and small language models in LDS tasks.
   The work provides empirical evidence that, while large language models achieve superior overall performance, small language models can offer competitive results in specific dimensions such as conciseness and efficiency, highlighting a trade-off relevant for practical deployments.
6. Validation through a large-scale real-world case study in oceanographic decision-making.
   The proposed framework is applied to a high-dimensional, real-world dataset (over 200 million data points per tree), demonstrating its capability to extract meaningful patterns and support decision-making in complex and uncertain environments.
7. Demonstration of decision-support capabilities in renewable energy contexts.
   The generated linguistic summaries enable the interpretation of geophysical dynamics and provide actionable insights for hydrokinetic energy assessment, illustrating the applicability of the approach in energy planning scenarios.

Overall, this work advances the state of the art in linguistic data summarization by introducing a probabilistically grounded and generatively enhanced methodology that improves scalability, interpretability, and practical relevance in data-intensive domains.

4.4. Limitations and Future Research Directions

Despite the promising results obtained, several limitations of the proposed framework must be acknowledged.

First, the computational cost associated with probabilistic structure learning and the integration of generative models remains significant. Although the approach is suitable for offline analysis of large datasets, its applicability to real-time or streaming environments is still limited.

Second, the framework assumes high-quality and preprocessed input data. In real-world scenarios, data may contain noise, missing values, or inconsistencies, which could affect the stability of the probabilistic models and, consequently, the quality of the generated linguistic summaries.

Third, while the use of GenAI enhances interpretability, it also introduces variability in the generated outputs. Although this variability was mitigated through statistical validation and repeated experiments, full determinism cannot be guaranteed, which may be critical in high-stakes decision-making contexts.

Finally, the validation of the approach was conducted within a specific domain (oceanographic data). Although the results are encouraging, further evaluation across different domains is required to assess the generalizability of the method.

Future research should focus on reducing computational complexity to enable near real-time applications, as well as on integrating robust data preprocessing and uncertainty quantification mechanisms. Additionally, exploring the incorporation of explainable AI (XAI) techniques could further enhance transparency and user trust in the generated summaries. Cross-domain validation and the development of domain-adaptive linguistic grammars also represent promising directions for extending the applicability of the proposed framework

5. Conclusions

This work demonstrates the power of hybridizing linguistic data summarization techniques, algorithms for constructing probabilistic trees (Chow-Liu), and different models of GenAI. This

 doi:10.20944/preprints202604.1304.v1

combination of techniques allowed for the processing of large volumes of data and the generation of linguistic summaries that facilitate data comprehension.

An important element is the treatment of information uncertainty and the triangulation of methods that enhance high consistency in the discovered knowledge. The generation of linguistic summaries incorporates into the proposed algorithms a high capacity for simulating human tolerance in the decision-making scenarios presented in the case studies. The proposed algorithms allowed for the processing of numerical, ordinal, and categorical variables. By employing a Controlled Natural Language (CNL) integrated with the GenAI agents, it was possible to "humanize" technical data, transforming it into easily interpretable linguistic protoforms to aid in decision-making.

However, the results obtained identify the computational cost as a limitation of the proposal, restricting the application of the algorithms to asynchronous use on historical data rather than real-time processing. A future line of research is identified as addressing the efficiency and use of real-time extensions of the proposed model.

The algorithms were validated in three experiments. The first experiment allowed for a comparison of the effectiveness of different GenAI models in generating linguistic summaries.

1. The main conclusion of Experiment 1 is that the results obtained by Gemini were significantly superior to the other two GenAI models used.
2. The GPT-5 LLM performed significantly better than the GPT-5 Mini on specific indicators of accuracy and structure, such as T3 and T4. However, in the overall analysis of the validation scenario, no significant differences were found between these two models.
3. The GPT-5 Mini model, by its very nature, generated significantly shorter or more concise summaries than the Gemini and GPT-5 models.
4. It is identified that there are specific dimensions of language processing where small models (SLMs) can be more effective or accurate than their LLM counterparts.

Finally, the validity of these results is supported by high agreement and statistical significance. Friedman tests and Kendall's W coefficient (with values up to 0.95) were used. Non-parametric tests and the Holm-Bonferroni correction reinforce the integrity of the model comparison study.

The decision-support capacity of the proposed algorithms is validated through their application in the study of oceanographic data, with measurements taken from satellites where uncertainty and imprecision are present.

The validation of the proposal through methodological triangulation techniques demonstrates that linguistic summaries are highly effective decision-making tools, successfully transforming 23 years of complex geostrophic variable data into readable information with high levels of confidence (frequently exceeding 80%). From an oceanographic perspective, the work identified.

The region exhibits a current structure where the anomaly (mesoscale/eddies) is the main driver of absolute velocity. Seasonality primarily affects sea level (adt/sla), but does not alter the strong dependence between velocity components and their respective anomalies.

A particularly high dependence on the zonal component (U) is noted, suggesting that coast-ocean exchange processes and coastal jets are the main energy vectors in the region.

Regarding temporal and climatic variability, the study concludes that the system shows marked sensitivity to large-scale events such as El Niño (ENSO). During these periods, as observed in 1997-1998 and 2015-2016, sea level and geostrophic velocity anomalies become the main drivers of the system, displacing the influence of permanent currents and increasing Geostrophic Kinetic Energy (KKE).

While winter offers a more stable and consistent scenario for oceanographic analysis, summer and years of climate transition present a more energetic but chaotic dynamic, characterized by greater data dispersion and an immediate response to remote forcings such as equatorial Kelvin waves.

Finally, the analysis of marine energy extraction potential reveals that the Norte Grande region possesses a significant hydrokinetic resource, although its exploitation requires technologies adapted to changing mesoscale flows rather than constant currents. Strategic "hot spots" are identified, such

 doi:10.20944/preprints202604.1304.v1

as the Mejillones Peninsula, Iquique, and the Arica Elbow, where the interaction of the current with the coastal topography maximizes energy density.

The study reveals that winter is the ideal period for stable electricity generation due to its high predictability, unlike summer, where energy peaks occur.

For citations of references, we prefer the use of square brackets and consecutive numbers. Citations using labels or the author/year convention are also acceptable. The following bibliography provides a sample reference list with entries for journal articles [1], an LNCS chapter, a book [3], proceedings without editors [4], as well as a URL [5].

Author Contributions: Conceptualization, IPP, LAA and PPP; Methodology, IPP, LAA and PPP; Software, IPP, PPP and RYH; Validation, IPP, P.P.P. and MLV; Formal Analysis, IPP and PPP; Investigation, IPP, LAA and PPP; Resources, IPP, LAA and RYH; Data Curation, IPP and PPP; Writing – Original Draft Preparation, IPP; Writing – Review & Editing, LAA, PPP and RYH; Visualization, IPP and MLV; Supervision, RYH; Project Administration, LAA. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author**.**

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

Acknowledgments

The authors would like to acknowledge the Copernicus Marine Service for providing open-access, high-quality oceanographic data essential for this study. We also thank the P.P. Shirshov Institute of Oceanology (IO RAS) for its contribution to oceanographic research and scientific knowledge that supports studies of this nature.

During the development of this research, GenAI tools, including GPT-5, GPT-5 Mini, and Gemini Flash 2.5, were used for the generation and evaluation of linguistic summaries. Small language models were deployed locally using OpenWebUI v0.6. All outputs generated by these tools were carefully reviewed, validated, and interpreted by the authors. The authors take full responsibility for the content of this publication.

References

1. Pérez Pupo, I., Piñero Pérez, P.Y., Bello Pérez, R.E., García Vacacela, R., Villavicencio Bermúdez, N.: Linguistic Data Summarization: A Systematic Review. In: Piñero Pérez, P.Y., Bello Pérez, R.E., Kacprzyk, J. (eds.) Artificial Intelligence in Project Management and Making Decisions, pp. 3–21. Springer, Cham 2022. DOI: 10.1007/978-3-030-97269-1_1
2. Phong PD, Lan PT, Thanh TX.: Linguistic Summarization and Outlier Detection of Blended Learning Data. Applied Sciences, vol. 15, no. 12 2025. DOI: 10.3390/app1512664
3. Tran, X.T.; Pham, D.P.; Pham, T.L. A Novel Linguistic Summarization of Time Series Data Based on Enlarged Hedge Algebra Formalism and Genetic Algorithm. Indones. J. Electr. Eng. Inform. 2026, 14, 281–292. DOI: 10.52549/ijeei.v14i1.6877
4. Veens, M.M.A. The use of linguistic summarization in a clinical protocol improvement context: a case study on a ICU glucose control protocol. Master Thesis, Eindhoven University of Technology, Eindhoven, January 2021.
5. Alvey B, Anderson D, Keller J.: Minimizing Protoform Redundancy to Enhance Linguistic Summaries in Object Detection," in 2025 IEEE International Conference on Fuzzy Systems (FUZZ), pp. 1–8, 2025. DOI: 10.1109/FUZZ62266.2025.11152168
6. Ahmed, M.: Data Summarization: A Survey. Knowledge and Information Systems 58(2), 249–273, 2019. DOI: 10.1007/s10115-018-1183-0

 doi:10.20944/preprints202604.1304.v1

7. Pérez Pupo, I.: Algoritmos para la sumarización lingüística de datos, aplicaciones en la toma de decisiones en la gestión de proyectos. Doctoral dissertation, Universidad de las Ciencias Informáticas, Cuba, 2021.
8. Gupta N, Singh V, Iyer A, Shiragur K, Grover P, Bairi RB, Maiti R, Damle S, Gupta SM, Maurya R, C VD.: Chow-Liu Ordering for Long-Context Reasoning in Chain-of-Agents. arXiv, 2026. DOI: 10.48550/arXiv.2603.09835
9. Kaczmarek-Majer, K., Baczyński, M., Hryniewicz, O., Miś, K., Mucha, W., Wichrowski, F.: Fuzzy Linguistic Summaries for Hidden Markov Models. In: Information Processing and Management of Uncertainty in Knowledge-Based Systems, pp. 266–276, 2025. DOI: 10.1007/978-3-031-73997-2_23
10. Kaczmarek-Majer, K., Casalino, G., Castellano, G., Leite, D., Hryniewicz, O.: Fuzzy Linguistic Summaries for Explaining Online Semi-Supervised Learning. In: 2022 IEEE 11th International Conference on Intelligent Systems (IS), pp. 1–8, 2022. DOI: 10.1109/IS57118.2022.10019636
11. Boix-Adserà E, Bresler G, Koehler F.: Chow-Liu++: Optimal Prediction-Centric Learning of Tree Ising Models. In 2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS), pp. 417–426, 2022. DOI: 10.1109/FOCS52979.2021.00049
12. Lou X, Hu Y, Li X.: Learning Linear Polytree Structural Equation Model. Transactions on Machine Learning Research, 2025
13. Rebane, G., Pearl, J.: The Recovery of Causal Poly-Trees from Statistical Data. In: Third Conference on Uncertainty in Artificial Intelligence, 2013. DOI: 10.48550/arXiv.1304.2736
14. Minh DLP.: Bayesian networks: Exact inference via macro-node polytrees. Communications in Statistics - Theory and Methods, vol. 0, no. 0, pp. 1–46, 2026. DOI: 10.1080/03610926.2025.2602732
15. Kacprzyk, J., Yager, R.R.: Linguistic Summaries of Data Using Fuzzy Logic. International Journal of General Systems 30(2), 133–154, 2001. DOI: 10.1080/03081070108960702
16. Copernicus: Global Ocean Gridded L4 Sea Surface Heights and Derived Variables Reprocessed (1993–ongoing). Available at: https://data.marine.copernicus.eu/product/SEALEVEL_GLO_PHY_L4_MY_008_047/description (2024). Accessed 29 Jan 2026. DOI (product): https://doi.org/10.48670/moi-00148