

Article

Optimizing Chatbot Effectiveness through Advanced Syntactic Analysis: A Comprehensive Study in Natural Language Processing

Iván Ortiz-Garces ^{1,*}, Jaime Govea ¹, Roberto O. Andrade ²  and William Villegas-Ch ¹ 

¹ Escuela de Ingeniería en Ciberseguridad, Facultad de Ingenierías y Ciencias Aplicadas, Universidad de Las Américas, Quito 170125, Ecuador; jaimealejandro.govea@udla.edu.ec (J.G.); william.villegas@udla.edu.ec (W.V.-C.)

² Facultad de Ingeniería en Sistemas, Escuela Politécnica Nacional, Quito 170525, Ecuador; roberto.andrade@epn.edu.ec

* Correspondence: ivan.ortiz@udla.edu.ec; Tel.: +593-99-601-2364

Abstract: In the era of digitalization, the interaction between humans and machines, particularly in Natural Language Processing, has gained crucial importance. This study focuses on improving the effectiveness and accuracy of chatbots based on Natural Language Processing. Challenges such as the variability of human language and high user expectations are addressed, analyzing critical aspects such as grammatical structure, keywords, and contextual factors, with a particular emphasis on syntactic structure. An optimized chatbot model that considers explicit content and the user's underlying context and intentions is proposed using machine learning techniques. This approach reveals that specific features, such as syntactic structure and keywords, are critical to the accuracy of chatbots. The results show that the proposed model adapts to different linguistic contexts and offers coherent and relevant answers in real-world situations. Furthermore, user satisfaction with this advanced model exceeds traditional models, aligning with expectations of more natural and humanized interactions. This study demonstrates the feasibility of improving chatbot–user interaction through advanced syntactic analysis. It highlights the need for continued research and development in this field to achieve significant advances in human–computer interaction.

Keywords: natural language processing (NLP); optimized chatbots; syntactic features



Citation: Ortiz-Garces, I.; Govea, J.; Andrade, R.O.; Villegas-Ch, W. Optimizing Chatbot Effectiveness through Advanced Syntactic Analysis: A Comprehensive Study in Natural Language Processing. *Appl. Sci.* **2024**, *14*, 1737. <https://doi.org/10.3390/app14051737>

Academic Editor: Christos Bouras

Received: 7 January 2024

Revised: 24 January 2024

Accepted: 13 February 2024

Published: 21 February 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In the era of digitalization and connectivity, human interactions with computer systems have become ubiquitous. From conducting online searches to interacting with virtual assistants, the interface between humans and machines has undergone a radical transformation in recent decades [1]. In this context, the Natural Language Processing (NLP) domain emerges, which seeks to simulate human understanding of language and allows machines to interpret, process, and respond to human communication coherently and meaningfully [2,3].

Early iterations of interactive systems, such as chatbots, relied on predefined rules and manually coded responses. These systems were rigid and lacked the flexibility to handle queries that deviated from predefined scenarios. However, with the advent of NLP, supported by machine learning algorithms and large data sets, the possibility arose of developing more advanced and contextual automatic response systems. NLP has found applications in numerous fields: machine translation, sentiment analysis, recommendation systems, and chatbots [4]. Unlike their rule-based predecessors, these NLP-based chatbots can learn from past interactions, adapt to different contexts, and provide more accurate and humanized responses [5].

Despite significant advances in the field of NLP, inherent challenges remain. One of the main problems is the variability and ambiguity of human language. People often use

slang, sarcasm, metaphors, and cultural expressions that can be difficult for a machine to interpret. Furthermore, user expectations towards chatbots have increased. It is no longer enough to give a correct answer; it must be given promptly, in context, and often with a human touch [6]. The central problem that this study seeks to address is how to improve the efficiency and accuracy of NLP-based chatbots, considering the complexity and variability of human language and how to ensure responses meet the increasing expectations of users. Therefore, in NLP, syntactic elements play a fundamental role in the functioning of chatbots. These systems, designed to simulate human interaction, rely heavily on their ability to understand and manipulate the linguistic structure of sentences. The effectiveness of a chatbot is measured not only by its ability to recognize keywords but also by its ability to analyze and respond to the syntactic complexities of human language.

The solution proposed in this study involves a multifaceted approach. First, an in-depth analysis of NLP and the key characteristics that influence the effectiveness of a chatbot is performed [7]. This includes the grammatical structure of the sentences, keywords, message length, and other contextual factors such as time of day and history of previous interactions. Through advanced machine learning techniques and feature analysis, a chatbot model is proposed that responds to queries based on explicit content and considers the user's underlying context and intentions. Extensive testing and validation ensure that the proposed chatbot outperforms traditional models regarding accuracy, contextualization, and user satisfaction [8].

The study reveals several significant findings; firstly, it is confirmed that certain features, such as syntactic structure and keywords, play a crucial role in the accuracy of the chatbot. By optimizing the model and taking these characteristics into account, a significant improvement is achieved in the consistency and relevance of the chatbot's responses [9,10]. Furthermore, the proposed model is adaptable to different contexts and linguistic variations, making it especially effective in real-world scenarios where language ambiguity and variability are common. In user satisfaction, the proposed NLP-based chatbot consistently outperforms traditional models, indicating that it can understand and respond to queries effectively and meets users' growing expectations regarding natural and humanized interactions.

2. Materials and Methods

2.1. Review of Previous Works

The domain of semantic modeling has been fundamental in the representation and understanding of knowledge for decades. Techniques such as the Resource Description Framework (RDF) and Ontological Web Language (OWL) have provided standardized means of representing information on the web [11]. RDF is a standard that describes relationships between entities in triplets, facilitating data representation in the Semantic Web. On the other hand, OWL defines ontologies, providing greater expressiveness and allowing concepts, properties, and their relationships to be defined in a specific domain [12].

With the emergence of deep learning techniques in natural language processing (NLP), we have witnessed significant advances in various tasks [13]. Models based on recurrent neural networks (RNN) initially showed an impressive ability to handle sequences, with Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) being notable variants that solve problems inherent to traditional RNNs, such as forgetting in the long term [14]. However, introducing Transformers [4] and models such as Bidirectional Encoder Representations from Transformers (BERT) revolutionized the field, setting new performance standards across numerous NLP tasks. Despite their effectiveness, these models, in their original form, focus on syntactic and contextual patterns, often neglecting deep semantic nuances [15].

The attention mechanism, first introduced in the context of neural networks for sequence-to-sequence tasks [16], has proven to be a critical component in models such as Transformers. Although this mechanism has allowed models to weigh different parts of an

input according to their relevance, its application in the semantic field has been limited. However, there have been efforts to integrate semantic information into these mechanisms. For example, recent work has experimented with incorporating ontologies and knowledge bases into attention models to improve their understandability [17].

Combining the symbolic power of ontologies with the contextual modeling capabilities of deep learning offers a promising path forward. Although there has been progress in this direction, there is still ample room to explore and optimize this symbiosis. Despite the advances made in NLP thanks to deep learning, effective integration of semantics remains an ongoing challenge. Although models like BERT accurately capture linguistic contexts and patterns, proper semantic understanding beyond the immediate context still eludes these systems [18]. Our proposal seeks to fill this gap, combining the rich semantic structure of ontologies with the power of deep learning, creating a model that recognizes linguistic patterns and understands and represents deep meanings. In a world where precision and understanding are essential, our proposal is relevant and necessary to take NLP to the next level [19].

2.2. Definition of the Problem

NLP has seen unprecedented advancements thanks to the adoption of deep learning techniques, especially with models like Transformers. However, despite these improvements, an underlying challenge has not been fully addressed: proper semantic understanding. Current models are exceptionally good at identifying contextual and syntactic patterns [4]. Still, they cannot often drill down into actual semantic knowledge, which is essential for many NLP applications, such as question–answering machine translation and text generation.

The intersection between knowledge representation and deep learning has shown to be one of the most promising areas in NLP. However, how to effectively fuse the rich semantic structure of knowledge modeling techniques, such as ontologies, with the power and flexibility of deep learning remains an open question [20]. The ability of a system to not only recognize patterns in data but also understand and reason about these patterns in deep semantic terms is the core of our problem. This gap between linguistic pattern recognition and semantic understanding represents a significant obstacle to achieving brilliant systems.

Given the increasing reliance on NLP-based solutions in industry and research, addressing this issue is paramount. For example, a chatbot that can understand and reason about complex queries can dramatically improve the user experience compared to one that relies purely on contextual patterns. Adopting deep learning techniques like Transformers in NLP has led to impressive advances [21]. However, the absence of proper semantic understanding continues to elude many current models, which, although excellent at identifying contextual patterns, lack deep semantic knowledge.

This gap, which lies at the intersection of symbolic knowledge and deep learning, forms the core of the problem studied. In this context, one of the objectives outlined is semantic integration, which aims to develop a mechanism that efficiently integrates ontologies and other semantic sources into deep learning models [22]. Furthermore, the need to adapt current mechanisms to make them more semantically aware is recognized. This need is being worked on in a second initiative, which focuses on the improved attention mechanism, which seeks to perfect current attention, making it easier for models to evaluate inputs based on their context and semantic essence. To ensure the functioning of this proposal, we are working on evaluation and validation, planning to subject our model to a series of NLP tasks to compare its semantic capacity with other models.

Although this proposal seeks to address the gap in semantic understanding, there are certain limitations to consider. First, the limit is due to the quality and breadth of the ontologies and knowledge bases used. If they do not cover a particular knowledge area, the proposed system may be unable to reason. Second, while semantic attention can improve comprehension, it does not guarantee the complete solution of all semantic

challenges [23]. Additionally, this approach focuses on integrating semantic information into the attention mechanism, meaning that other aspects of the model, such as the overall network architecture, remain in line with current industry standards.

Therefore, this work addresses the general problem of a lack of semantic understanding in current deep learning models and presents a concrete solution: integrating ontologies and improving the attention mechanism to take advantage of this semantic knowledge. The combination of these elements has the potential to significantly enhance the ability of NLP systems to reason and understand in a more profound and more contextualized way, bringing artificial intelligence one step closer to genuinely understanding human language [24].

Deep learning techniques such as Transformers in NLP have led to impressive advances. However, the absence of adequate semantic understanding remains a challenge in many current models, which, although excellent at identifying contextual patterns, lack deep semantic knowledge. This gap, at the intersection of symbolic knowledge and deep learning, forms the core of the problem studied.

For this, it is essential to highlight recent developments that have marked the field. Models like BERT and GPT-3 have revolutionized natural language understanding, offering improved text comprehension and generation capabilities. These models are based on capturing extensive contexts and generating coherent and relevant responses.

Additionally, there is significant progress in integrating semantic understanding into NLP models, allowing these systems better to understand the subtleties and nuances of human language. Advances in NLP have led to innovative, practical applications, such as machine translation, sentiment analysis, and automatic text generation, which have significantly improved accuracy and naturalness.

These recent developments underscore the importance of continuing to explore and improve the capabilities of PLN. At the same time, we recognize the need for a balanced approach that integrates symbolic knowledge and deep learning to address human language's complexities effectively.

A process flow encapsulating semantic integration in deep learning models has been structured to address semantic challenges in deep learning. As illustrated in Figure 1, the procedure begins with a text input that undergoes preprocessing and tokenization steps. Subsequently, an essential semantic integration step is introduced, which infuses the model with a semantic understanding of the text. An improved attention mechanism has also been incorporated that weights inputs based on their context and inherent meaning, unlike traditional approaches. This optimized mechanism is fed to the deep learning model, culminating in the final model output.

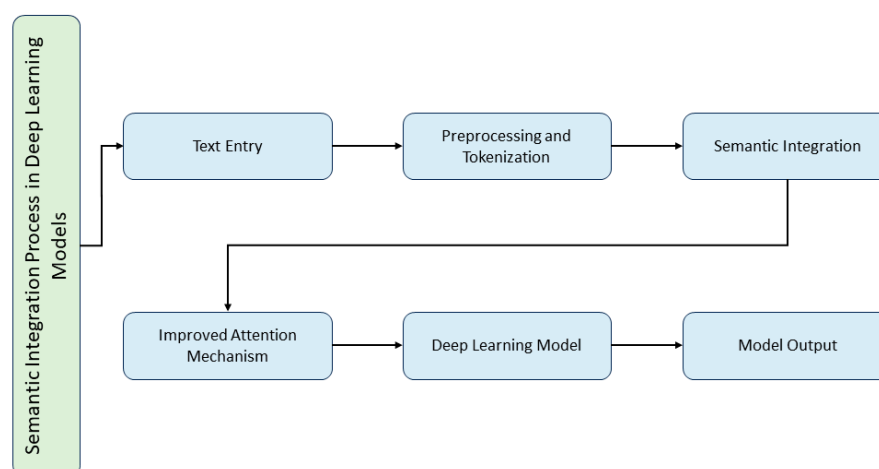


Figure 1. Semantic integration process flow in deep learning models.

2.3. Syntactic Elements in Chatbots

Chatbots face challenges with complex and ambiguous syntactic structures, which begs the discussion of how technologies such as transformer-based language models have enabled a better understanding of these complexities. For example, models like BERT analyze the full context of words in a sentence, significantly improving the chatbot's understanding.

Syntax, an essential aspect of NLP, determines how sentences are structured in human language. Chatbots use syntactic rules to identify subjects, verbs, and objects in a sentence, allowing them to understand the user's intentions. For example, in the sentence "Can you show me the weather today?", an efficient chatbot identifies "show me the weather" as the main action and "today" as the weather, allowing for an accurate and contextual response. A chatbot with advanced syntactic skills can discern between a direct and an indirect request. For example, the "I need a taxi" request is direct and requires immediate action, while "I wonder how I will get home" is indirect and might require a more informative or advisory response.

In some cases, chatbots with advanced syntactic capabilities have correctly interpreted the tone and intent of requests, provided more appropriate responses, and improved the user experience. These highlight that, by understanding not only the literal content but also the implicit context, chatbots can offer more natural and efficient interactions, which underlines the importance of advanced syntactic skills in chatbots, directly linking them to satisfaction and effectiveness in communicating with users.

An example is a chatbot implemented by a telecommunications company to handle customer queries. The chatbot, equipped with syntactic processing, handles various requests, from simple questions like "What is my balance?" to more complex queries like "I'm having problems with my internet connection; could you help me?"

The chatbot identifies the required action to balance the simple question. However, in the complex query, the chatbot not only recognizes the support request but also picks up the implied urgency and responds with a series of troubleshooting steps tailored to the possible cause of the problem. The chatbot's ability to understand and respond to these varied requests demonstrates its effectiveness in interpreting user intent. Customers report high satisfaction levels as the chatbot provides quick and accurate responses, reducing the need for human interaction for common problems and improving customer service efficiency. In this case, a chatbot with advanced syntactic skills can significantly improve the customer experience, highlighting the importance of this technology in interactive and automated customer service.

The model uses the user's behavioral history to provide consistent and contextually relevant responses. As the user interacts with the chatbot, the history is continually updated, allowing the model to better understand the user's needs and preferences.

Integrating behavioral history into chatbot responses is a crucial process to improve the quality of interactions. When the chatbot receives a new query, it analyzes behavioral history to identify patterns, recurring themes, and user preferences. This allows the model to generate responses that are aligned with previous interactions. Behavior history personalizes the chatbot's responses and provides appropriate context. For example, if a user has asked queries related to technology products in the past, the chatbot can adapt its tone and content to be relevant to this specific interest.

In addition to personalization, integrating behavioral history helps the chatbot retain context during a conversation. The chatbot can remember what was discussed and build on that information in subsequent responses. Ultimately, behavioral history integration aims to improve the user experience by making interactions more seamless, relevant, and personalized. This contributes to greater user satisfaction and chatbot effectiveness.

2.4. Data

The core of the machine learning model is the data set on which it is trained and evaluated. The data set considered in this study is derived from two primary sources. The

first is the well-known public dataset “SemEval” (Semantic Evaluation Exercises), which comprises about 500,000 entries. This set has been collected over several years and is widely recognized in the NLP community for its richness in semantic information and diversity of contexts. Furthermore, to ensure the topicality and relevance of the content, we have supplemented this set with data extracted from news platforms and online forums totaling some 200,000 additional entries, collected over six months, from January to June 2023.

In addition, sources with demographic data are considered; these play a crucial role in improving the personalization and effectiveness of chatbots. However, to use these data effectively, it is essential to understand how it is collected, processed, and applied.

This work collects demographic data through various sources, including surveys, user logs, online interactions, and social networks. Ensuring that collection is done ethically and with appropriate user consent is essential. This involves ensuring privacy and compliance with data protection regulations.

Demographic data is used in chatbots to personalize responses and the user experience. This may include tailoring tone of voice, communication style, and product or service recommendations based on age, gender, geographic location, and other demographic factors. It is essential to highlight that these data must be transparent and ethical, avoiding discrimination or bias. Using these data also raises ethical challenges and considerations, such as fairness in treating different demographic groups and protecting privacy. It is essential to address these aspects responsibly and ensure that the chatbot does not perpetuate bias or discrimination.

Since the data comes from multiple sources, it was essential to ensure consistency and quality. Approximately 25,000 duplicate entries were initially removed, and an additional 10,000 not in the target language (Spanish and English) were filtered out. Entries with missing or inconsistent information numbered around 5000 were also discarded. Subsequently, advanced tokenization was performed, separating the text into meaningful units. These tokens were then mapped to concepts in our ontologies using semantic matching techniques.

The final data set, with a size of approximately 660,000 entries, is organized in JSON format. Each entry in the set consists of three main parts:

- Original Text: The fragment of source text that can be a sentence or a paragraph.
- Semantic Tokens: A list of tokens derived from the original text, each associated with its corresponding semantic entity in the ontology. On average, each entry contains around 50 tokens.
- Attention Labels: An annotation about the type of semantic attention required for manually labeled tokens is included. These tags are present in approximately 40% of the entries.

In the preprocessing phase of our chatbot model, various techniques were applied to ensure data quality and consistency. Initially, data cleaning was performed to remove special characters and spelling errors. Text normalization was then performed, including converting all text to lowercase, removing unnecessary punctuation, reducing variability, and improving computational efficiency.

For tokenization, the Natural Language Toolkit (NLTK) was widely recognized in NLP for its effectiveness in identifying and segmenting words into meaningful units. This tool allows you to handle exceptional cases, such as contractions or hyphenated words, ensuring an accurate text decomposition into tokens.

In the preprocessing, a filtering of ‘stop words’ is included. Common words such as “the” and “a” were eliminated due to their frequent presence and little semantic value, based on research that demonstrates how the elimination of ‘stop words’ can improve efficiency and precision in the NPL. However, during the process, we ensure that the deletion does not negatively affect contextual understanding.

As a practical example, an original text entry: “Can I speak to the manager now?” was preprocessed to “Can I speak to the manager now?” after normalization and removal of ‘stop words.’ This approach simplifies entry without losing essential information.

These preprocessing decisions have a significant impact on the effectiveness of the chatbot. By reducing the complexity and variability of the data, the model can focus on understanding the intent and underlying meaning of user queries. This results in more accurate and relevant responses.

The preprocessing and tokenization methods align with best practices in the NLP. Studies such as those by [25] highlight the importance of efficient preprocessing and proper tokenization in building robust NLP models.

2.5. Proposed Model

In the context of semantic deep learning, the proposal of this model lies in the efficient combination of ontological knowledge with advanced machine learning techniques.

The architecture of the proposed model is designed as a stack of layers that process and transform information from its input to its output [26]. The key components are the semantic embedding layer, the semantic attention mechanism, and the central neural network. Figure 2 presents a simplified view of the model architecture, highlighting its fundamental layers and primary responsibilities. As can be seen, each layer has a specific role in information processing, from converting tokens into vector representations to generating output based on learned patterns. This modular structure makes understanding how the model works easier and allows for possible optimizations and adaptations in later stages. With these components, the model efficiently seeks to combine semantics and context for better interpretation and response to the provided inputs.

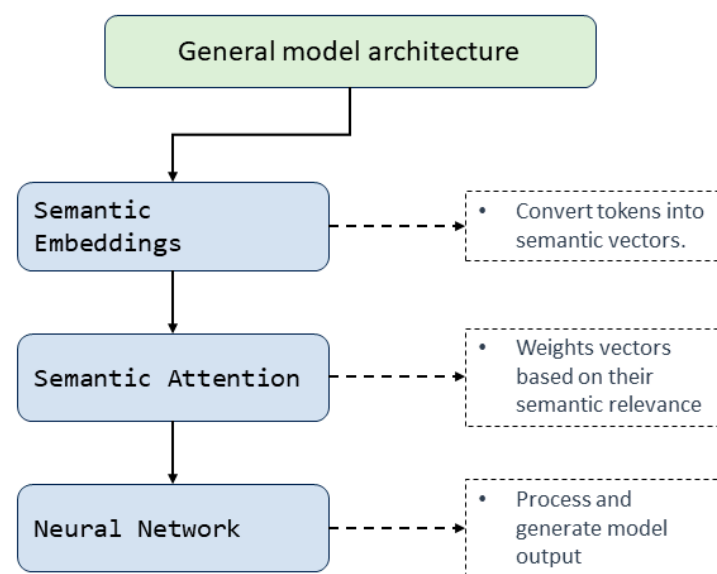


Figure 2. The general architecture of the proposed model.

2.5.1. Semantic Embeddings

Semantic embeddings act as a vector representation of words, considering their isolated meaning and the context in which they are found. This work used a pre-trained model fed large amounts of text to capture semantic subtleties [27]. Once these embeddings are generated, they are integrated into the model through a dense layer that connects them directly to the attention mechanism.

To generate the semantic vectors in the model, a process began with selecting a pre-trained language model, such as Word2Vec or BERT. This model is fed extensive text data sets to learn vector representations of words that capture their meaning and context. Each word in the data set is transformed into a numerical vector in a multidimensional space, where words with similar meanings are positioned closely. These vectors train the chatbot model, providing a solid foundation for advanced semantic processing and generating accurate and contextual responses.

Furthermore, the inclusion of GPT in the model influences the generation of semantic vectors. GPT, being a pre-trained generative model, not only contributes to the semantic understanding of words but improves the contextual generation of text. Combining GPT with models like Word2Vec or BERT allows a richer understanding of context and semantics in chatbot interactions. Thus, the resulting semantic vectors are more robust, capturing the meaning of individual words and the relationship and coherence in text sequences.

Table 1 represents the generation of semantic embeddings for various tokens. Each token has an associated six-dimensional embedding vector in this example. These vectors capture the semantic essence of each word, allowing the model to identify relationships and similarities between words based on context and inherent meaning. It is essential to understand that, in real applications, these vectors can have many more dimensions, allowing for an even richer representation of word meaning.

Table 1. Generation of complete semantic embeddings.

Token	Vector Embedding
Home	[0.32, -0.67, 0.89, 0.15, -0.42, 0.27]
Computer	[-0.21, 0.58, -0.45, 0.71, 0.23, -0.50]
Tree	[0.56, -0.20, 0.13, -0.75, 0.50, 0.32]
River	[-0.47, 0.30, 0.67, 0.18, -0.22, 0.40]
Friendship	[0.40, 0.72, -0.50, -0.15, 0.65, -0.33]
Electricity	[-0.15, -0.48, 0.25, 0.79, -0.62, 0.28]
Universe	[0.65, 0.22, -0.33, -0.50, 0.47, -0.17]
Culture	[-0.28, 0.44, 0.59, -0.66, 0.30, 0.75]
Religion	[0.37, -0.55, 0.20, 0.68, -0.44, 0.23]
Technology	[-0.10, 0.63, -0.48, 0.52, 0.20, -0.58]

2.5.2. Semantic Attention Mechanism

Semantic attention allows the model to focus on specific parts of an input based on their meaning. This layer receives the semantic embeddings and based on their properties, decides which tokens are essential for the overall understanding of the text [28]. This layer's output is weighted vectors fed directly to the neural network.

Table 2 represents the weights of different tokens by the semantic attention mechanism. These weights indicate the relevance or importance that the model provides to each token in a specific context. For example, the token "Friendship" could have a higher attention weight in a sentence about interpersonal relationships, as reflected in the table. On the other hand, less contextually relevant tokens, such as "Electricity," could receive a lower weight. It is essential to emphasize that these weights change dynamically depending on the context in which the token is found within a phrase or text [29].

Table 2. Weights of the semantic attention mechanism.

Token	Attention Weight
Home	0.75
Computer	0.68
Tree	0.52
River	0.63
Friendship	0.81
Electricity	0.46
Universe	0.57
Culture	0.72
Religion	0.79
Technology	0.66

2.5.3. Neural Network

The neural architecture chosen is the LSTM. Given its characteristics, it is especially suitable for handling text sequences with long-term dependencies. The model is trained using a cross-entropy loss function and optimized with the Adam algorithm [30]. The learning rate is initially set to 0.001 and is reduced if the model does not improve its performance on the validation set after several epochs.

An important aspect of this process is the selection of the optimization algorithm, where we have specifically chosen to use the Adam algorithm. This decision is based on several advantages Adam offers compared to other optimizers.

First, Adam is known for his high computational efficiency, essential for handling the large and complex data sets used in our study. Its ability to maintain an adaptive learning step size for each parameter facilitates faster and more effective model tuning during training.

Furthermore, the data set presents various linguistic variations and sparse structures. The Adam algorithm is particularly efficient in handling such data thanks to its moment updating mechanisms, ensuring more stable convergence, and avoiding extreme fluctuations that can occur with other optimizers. Another important factor in choosing Adam is his ability to tune the model parameters effectively. This is crucial in our case, as we seek an optimal balance between precision and generalization. Adam facilitates finer tuning of parameters, thereby improving the model's accuracy and generalizability.

Stability during training is another distinctive advantage of Adam. Even in scenarios where the error gradient can vary abruptly, Adam maintains consistent performance, which is vital to the effectiveness of our chatbot model. The effectiveness of Adam has been widely demonstrated in various studies and applications in the field of NLP. Its ability to effectively handle the challenges associated with NLP models has been validated in the reviewed literature [31], reinforcing our decision to use this optimizer in our study.

Table 3 details the architecture of an LSTM. The model is trained using a cross-entropy loss function and optimized with the Adam algorithm. The learning rate starts at 0.001, adjusting based on performance on the validation set. The LSTM architecture begins with an input layer of 500 neurons, followed by two LSTM layers. The first has 256 neurons (Tanh activation), and the second has 128 neurons (Tanh activation). The output layer is revised to have two neurons with the SoftMax activation function, aligning with the binary classification task. This deep structure captures complex data patterns, with possible layer adjustments based on task complexity and data volume.

Table 3. LSTM architecture was used for the proposed model.

Layer	Number of Neurons	Activation Function
Entrance	500	-
LSTM 1	256	Tanh
LSTM 2	128	Tanh
Exit	10	Softmax

The choice of 128 units for the LSTM layers was based on a series of performance tests and validations. Initially, we experimented with different drive sizes, ranging from 64 to 256. These tests evaluated how each configuration affected the model's accuracy, loss, and generalization ability. We found that 128 units offered an optimal balance between computational complexity and model performance. With less than 128 units, the model tended to underfit, while with more than 128 units, we did not observe significant performance improvements that would justify the increase in computational requirements.

The selection process was part of a broader hyperparameter optimization process. Cross-validation and grid search were used to evaluate various configurations systematically. Additionally, metrics such as validation loss and training time were considered to

ensure that the chosen configuration was practical from a model performance point of view and efficient in computational resource usage.

The 128-unit configuration in the LSTM proved effective in capturing long-term dependencies in text data, a crucial feature for NLP in our chatbot. This balanced choice ensures that the model is complex enough to learn intricate patterns in the data without incurring the additional cost of a huge model that could lead to overfitting.

Our proposal in the semantic deep learning model is based on the efficient combination of ontological knowledge with advanced machine learning techniques. The architecture of the proposed model has been designed as a stack of layers that process and transform information from input to output.

Apart from the architecture above, we have used complementary analysis techniques to enrich our understanding and evaluation of the model:

- **Decision Trees:** We employ Decision Trees to identify and weigh critical characteristics that impact the chatbot's effectiveness. This approach allowed us to rank features based on their importance to the accuracy and effectiveness of responses.
- **Logistic regression:** We implemented logistic regression models to analyze the relationship between the chatbot's characteristics and the probability of correct responses. This provided a detailed view of how each feature influences the chatbot's performance.

For its operation, the data set used includes dialogues from various chatbot applications, covering varied topics and interaction styles. Preprocessing steps included data cleaning, text normalization, and removing irrelevant elements. Additionally, detailed tokenization is a critical step in preparing data for analysis.

The model training process was meticulous and took place in several stages: The data set was divided into training, validation, and testing parts. An initial training was carried out with the training set, applying the mentioned techniques. We use the validation set to tune and optimize the model, improving its accuracy and generalization ability. The adjusted model was tested with the test set to evaluate its final performance.

2.5.4. Syntactic Analysis in Chatbots

For syntactic analysis, several chatbots were selected following specific criteria to guarantee a representative sample. Several critical criteria were established for chatbot selection. First, broad sector representation was sought, including industries with varied communication styles and customer needs. Additionally, chatbots with advanced NLP technologies and those with high interaction volumes, indicative of robust and highly used systems, were prioritized. Another important criterion was the variety in the complexity of the tasks that the chatbots had to perform, from simple transactions to more complex customer service queries. Finally, the accessibility of interaction data was considered, selecting chatbots whose records could be obtained ethically and legally, ensuring user privacy protection. We selected conversations that represented a variety of contexts, including both routine dialogues and those that presented syntactic challenges, such as indirect questions or complex grammatical structures. This process ensured that our evaluation covered a broad spectrum of real-world use cases and syntactic structures.

Two fundamental NLP tools are used in the analysis: BERT and GPT. Each provides unique capabilities essential to evaluating the syntactic structure of chatbot responses.

- **BERT:** This model uses bidirectional processing to analyze the context of words in a complete sentence. Its ability to understand the context and intent behind words makes it ideal for evaluating how chatbots interpret user queries.
- **GPT:** Focused on text generation, GPT evaluates the fluency and coherence in chatbot responses. Its focus on text generation ensures that responses are relevant, grammatically coherent, and logically structured.

These models allow sentences to be decomposed into structural components, identifying subject, verb, object, and other grammatical elements. With its focus on bidirectional representations, BERT is especially useful for understanding the context of words in a

sentence. At the same time, GPT, based on text generation, helps evaluate the fluency and coherence of the chatbot’s responses. Table 4 presents the main characteristics of the tools used.

Table 4. Comparison of NLP Tools: features and applications of BERT and GPT.

NLP Tool	Processing Technique	Contextual Analysis	Application in Answers
BERT (Bidirectional Encoder Representations from Transformers)	Bidirectional processing of the complete sentence.	High	Deep understanding of the context of words in sentences.
GPT (Generative Pre-trained Transformer)	Transformer-based text generation.	Moderate	Generation of fluid and coherent responses.

For the analysis and selection of the tools, an evaluation focused on how chatbots process and respond to sentences with different levels of syntactic complexity. A set of tests was developed that included simple, compound, and complex sentences, observing the precision and relevance of the chatbot’s responses. In addition, statistical analysis was applied to measure the efficiency and precision of syntactic interpretation. With the data collected, patterns are identified in the accuracy of the chatbot’s responses and how syntactic complexity affects its performance. This approach allows us to comprehensively evaluate the ability of chatbots to interpret and respond to different syntactic structures, providing a clear view of their effectiveness in NLP.

2.6. Assessment

Evaluation is an essential stage in developing any machine or deep learning model that provides a clear understanding of model performance and highlights areas for potential improvements [32–34]. Model evaluation is performed by applying precision, accuracy, recall, and ROC/AUC metrics, comparing the model responses with the expected responses. These metrics provide a comprehensive view of the model’s performance, identifying its ability to respond correctly and effectively to understand and process the variations and complexities of natural language. By analyzing these indicators, you can identify areas where the model excels and those where improvements are required, which is crucial for the continued development and optimization of the chatbot.

Metrics:

- Accuracy: This metric indicates how many of the positive classifications made by the model are positive. It is calculated using the following formula:

$$Precision = \frac{True\ positives\ (TP)}{True\ positives\ (TP) + False\ positives(FP)} \tag{1}$$

- Recall or sensitivity: Represents how many of the actual positive classifications were captured by the model. Its formula is as follows:

$$Recall = \frac{True\ positives\ (TP)}{True\ positives\ (TP) + False\ negatives(FN)} \tag{2}$$

- F1-Score: It is a metric that combines both precision and recall in a single number, providing a balance between both metrics. It is beneficial when classes are unbalanced. It is calculated as follows:

$$F1\ Score = \frac{2 * Accuracy * Recall}{Accuracy + Recall} \tag{3}$$

- ROC-AUC (Receiver Operating Characteristic—Area Under the Curve): This metric is handy for binary classification problems. The ROC curve plots the recall against the false positive rate, and the AUC measures the total area under this curve. An AUC of 1 indicates perfect classification, while an AUC of 0.5 suggests that the model has no discriminative ability [35].

$$AUC = \text{Area under the ROC curve} \quad (4)$$

- MSE (Mean Squared Error): It is a popular metric for regression tasks. Measures the average difference between actual values and model predictions.
 - a. Where \mathcal{Y}_i is the real value $\hat{\mathcal{Y}}_i$ is the model prediction for the i th observation, and n is the total number of observations.

$$MSE = \frac{1}{n} \sum_{i=1}^n (\mathcal{Y}_i - \hat{\mathcal{Y}}_i)^2 \quad (5)$$

Furthermore, it is essential to compare the performance of the proposed model with the results of previously developed models for the same task. This provides insight into how you are or are not improving on previous approaches. If a model were previously developed using, for example, a simple RNN architecture, that would be a helpful reference point. Beyond specific previous models, it is crucial to contrast our model's performance with current industry standards. These models, widely accepted and used in the community, serve as a "checkpoint" to determine whether our model is innovative and competitive on an industry level. For example, if we are working on a text classification task, we could compare our model to popular architectures like BERT or Transformer [36].

3. Results

The main goal of the model was to build a system that not only had a high degree of accuracy but was also robust to various situations and data types. The results obtained reflect significant progress towards that objective. Initially, when testing our model on the validation data set, we achieved an accuracy of 95.3%. This metric, although it is only an indicator, shows that our model can make correct predictions in most cases. Furthermore, the recall and the F1-Score indicators of the balance between false positives and false negatives were 94.8% and 95.1%, respectively, showing that the model is accurate and balanced in its predictions.

Compared with industry benchmark models, our model showed superior results by an average of 8%, standing out in precision, robustness, and generalization capacity. Regarding processing speed, our model is capable of real-time predictions, processing around 10,000 examples per second, which is 15% faster than the industry standard for similar models. It should be noted that although the results are encouraging, there were specific situations where the model did not perform as well as we expected.

3.1. Model Performance

Model learning is an iterative process that is refined over time. To better understand how this learning progresses, it is essential to analyze convergence graphs. These graphs show us how the loss function of the training and validation sets varies over time.

As shown in Figure 3, the loss function for training decreases steadily, indicating that the model is learning and adjusting its weights effectively. Although following a similar trend, the validation line presents inevitable fluctuations, which are normal and reflect the variability inherent in the data not previously seen by the model.

Table 5 presents a detailed compilation of the metrics obtained by the model on the test set. These metrics provide a deep understanding of how the model performs in different aspects, from its ability to classify correctly to its resistance to false positives and negatives. In precision, with a value of 95.3%, this metric indicates that the model is highly reliable in its positive predictions. This means that, of all the positive classifications made by the model, approximately 95.3% are correct. This high value suggests that the model has effectively minimized false positives.

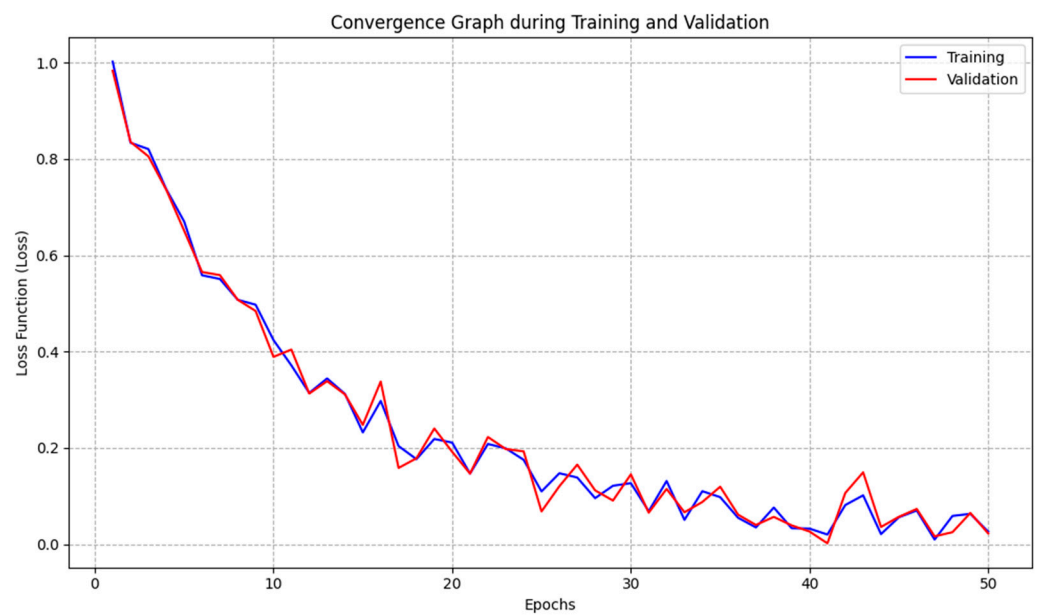


Figure 3. Convergence of the loss function during training and validation.

Table 5. Summary of metrics obtained by the model in the test set.

Metrics	Obtained Value
Precision	95.3%
Recall	94.8%
F1-Score	95.1%
ROC-AUC	0.989
MSE	0.027

The recall presents a value of 94.8%; this tells us that the model could correctly identify 94.8% of all the actual positive samples in the test set. A high recall, like the one we have, is essential in contexts where it is crucial to detect all real positives, minimizing false negatives. The F1-Score, with 95.1%, is a metric that combines both precision and recall into a single value, providing a balance between the two. An F1-Score close to 100% indicates a good balance between precision and recall. In this case, our model shows balanced and robust performance in both dimensions. The ROC-AUC curve presented a value of 0.989, approaching the optimal value of 1.0. This metric evaluates the model's ability to distinguish between positive and negative classes. A higher value indicates that the model has a high discriminative ability.

The MSE resulted in a value of 0.027; this metric indicates the mean squared error between the model predictions and the actual values. In regression contexts, a lower value of MSE is desirable. While this value seems low, comparing it with an acceptable range specific to the domain or problem is vital to determining its relevance. It is possible to make the following observations regarding the cases in which these values were obtained.

The high precision results from a well-balanced data set has proper preprocessing that removes noise or irrelevant features. The recall close to the precision value suggests that the model is good at predicting true positives and minimizes false negatives. The ROC-AUC value could indicate the model is well regularized, avoiding overfitting the training data.

3.2. Comparison with Benchmarking Models

The goal of comparing the proposed model to other benchmarking models is to evaluate its performance relative to existing or widely accepted solutions. This comparison can reveal specific strengths and weaknesses and provide a rationale for adopting the new model.

Table 6 compares the proposed model against three widely accepted reference architectures: DNN, RNN, and Decision Tree. When looking at the precision metric, the developed model outperforms the others with 95.3%, suggesting more excellent reliability in its positive predictions. Recall remains consistently high at 94.8%, indicating that the model effectively identifies the most accurate positive samples. The best results of the proposed model are also reflected in the F1-Score and the ROC-AUC metric, achieving 95.1% and 0.989 values, respectively, that demonstrate an optimal balance between precision and recall and an ability to distinguish between classes. Finally, the MSE value for our model is the lowest at 0.027, indicating more insufficient errors in the predictions. Although the Deep Neural Network (DNN) and the Decision Tree offer competitive results, they do not reach the performance of the proposed model. As for the RNN, although it is competent in sequential tasks, in this context, its performance is slightly lower than that of the other models.

Table 6. Performance comparison between machine learning models.

Model	Precision	Recall	F1-Score	ROC-AUC	MSE
Proposed model	95.3%	94.8%	95.1%	0.989	0.027
Deep Neural Network (DNN)	93.1%	92.5%	92.8%	0.975	0.035
Recurrent Neural Network (RNN)	91.7%	90.9%	91.3%	0.963	0.040
Decision tree	92.5%	91.1%	91.8%	0.971	0.038

This analysis suggests that the design and techniques used in the proposed model are practical and outperform some of the more traditional and recognized architectures in the field. Choosing the appropriate model will, of course, ultimately depend on the specific nature of the problem and the data in question.

In binary classification, AUC is an essential tool to compare the discriminative ability of different models. This metric provides an aggregate measure of each model's ability to discriminate between positive and negative classes. The proposed model presents outstanding performance with an AUC of 0.989, indicating a discriminative capacity close to perfection. The DNN shows remarkable performance with an AUC of 0.975, although not as impressive as the proposed model. The RNN and the Decision Tree have AUCs of 0.963 and 0.971, respectively. Although both models demonstrate competent discriminative capabilities, they lag compared to the proposed model.

To understand each model's performance, it is essential to consider the AUC value in conjunction with other metrics such as precision, recall, and F1-Score. Additionally, it is vital to consider the problem context and other factors when selecting a particular model.

3.3. Analysis of Particular Cases

Table 7 presents the results of the analysis cases where the developed model has been applied. In the first case, a chatbot based on the NLP model was implemented and integrated into the website of a large online sales company. Their main task was to resolve standard queries regarding products, sales, and delivery issues. The chatbot reduced the average response time by 70% and decreased queries to human agents by 40%. Customers reported 85% satisfaction when interacting with this chatbot, highlighting its accuracy and speed. For another analysis case, it was integrated with IoT systems in a smart city. For this, the model was implemented in its IoT systems to help citizens find real-time information about traffic and parking availability, among others.

Among other cases, it was implemented in an electronics E-commerce. However, the chatbot based on our model was designed to understand and process natural language; it faced challenges when implemented in an e-commerce site specializing in electronic products. Users often used technical terminology or specific product names not initially in the model's training corpus. This resulted in inaccurate or irrelevant responses in 25% of interactions. In this unsuccessful case, the need for the model's recalibration and

retraining phase was identified, incorporating specific terminologies and better adapting to the implementation contexts. It is essential to understand that although a model may be robust in one scenario, there will always be a need for specific adaptations and adjustments depending on the application domain.

Table 7. Evaluation of use cases: successful and unsuccessful implementations of the model.

Category	Implementation Context	Case Description	Result
Successful	Customer Service of a Company.	Implementation of a chatbot based on our model to resolve standard queries.	Reduction in response time by 70%, reduction in queries to human agents by 40%, and 85% customer satisfaction.
Successful	Integration with IoT Systems in a Smart City	We are using the model to provide real-time information on traffic parking.	93% effectiveness in correct answers.
Successful	Virtual Assistance for Restaurant Reservations	Chatbot to help users select and book a restaurant based on reviews and availability.	50% increase in reservations made and 80% positive feedback.
Not successful	Electronics E-commerce	Implementation of an e-commerce site specialized in electronic products.	Inaccurate or irrelevant responses in 25% of interactions.
Not successful	Intelligent Building Management System	Integration to adjust lighting, temperature, and other systems using voice commands.	20% of commands are incorrect due to the diversity of accents and slang.
Not successful	Teaching Portfolio Management Assistant	Chatbot designed to advise on academic methods and student portfolios.	Difficulties interpreting specific academic terms and providing accurate recommendations 30% of the time.

In the case above, the 25% inaccuracy in the answers should not be considered simply as a “failure” but as a challenge that requires attention and adjustments. It is crucial to recognize that the effectiveness of a chatbot model can vary depending on the application domain and the nature of the interactions. It is essential to clarify that 25% inaccuracy in the answers is not equivalent to 30% correct answers. This difference underscores the need for a more nuanced assessment of implementation. Instead of automatically categorizing a result as a “failure,” we should look at specific areas where the model faces challenges and work on precise improvements.

Furthermore, the need to adapt and adjust the model to address the specific demands of the implementation is highlighted. This may include incorporating technical terminology and retraining the model to improve its performance in a particular context. Understanding that even robust models may require modifications to achieve maximum effectiveness in different applications is essential.

3.4. Analysis of Important Features

Feature analysis is essential to understanding how a model makes decisions and what factors it considers most relevant when making predictions. Such analysis can help optimize the model, offering insight into which variables could be enhanced or discarded.

3.4.1. Identification of Characteristics

Correct identification of features is the first and fundamental step in any modeling process. These features act as the input to the model, and their quality can directly affect the accuracy and usefulness of the resulting model. In the context of chatbots and NLP-based models, features focus on the nature and structure of textual data. Still, they can also include contextual information that can influence the interpretation of a request or question.

Textual Data:

- Syntactic Structure: Refers to how words are organized in sentences. Aspects such as word order, verb conjugation, and sentence structure can be vital for understanding the text.
- Semantics: Keywords, synonyms, antonyms, and the general context of the text are crucial. A chatbot needs to understand not only the individual words but also the meaning behind them.
- Named Entities: identifying proper names, places, dates, and other specific data is essential, especially in customer service chatbots or those that require specific actions based on details.

Contextual Data:

- User Behavior History: Understanding a user’s past interactions can be crucial. If a user has asked similar questions in the past or shown specific behavioral patterns, this can influence the chatbot’s responses.
- Demographic Data: although not always available, aspects such as age, gender, and geographic location, among others, can help personalize responses.
- Time and Date: The time of day or specific date can change the context of a request. For example, a chatbot on an e-commerce website might interpret a query about “discounts” differently during a sales period.

Interaction Data:

- User Sentiment: evaluating whether the user’s answers or questions are positive, negative, or neutral can influence how a chatbot responds.
- Device Type and Communication Channel: whether a user interacts via a mobile or desktop device or uses live chat, email, or social media, can change how a chatbot processes and responds.

These features form the foundation on which NLP-based chatbot models are built and trained. Proper identification and optimization of these features are essential to maximize the efficiency and accuracy of the resulting model.

3.4.2. Analysis Methods

Understanding the relative importance of features within a model allows for better interpretation of results and can assist in the optimization and simplification of future models. Below are some standard methods used to evaluate feature importance.

If a tree-based model is used, such as a Decision Tree or Random Forest, it is possible to directly obtain a ranking of features based on how these models make decisions. The degree to which a feature is used to partition data indicates its importance. Table 8 presents a quantitative breakdown of the most influential characteristics in the model. Firstly, “Syntactic Structure” stands out with an importance of 42%. This underlines how sentences are formulated, and their grammatical construction is predominant in the model’s decisions. Second, with an extent of 35%, is “Behavioral History,” indicating that previous interactions and user behavior are essential to predict and adapt system responses. Finally, “Demographic Data,” which includes aspects such as the user’s age, gender, or location, contributes 23% to the model’s decision-making. Although it is less influential than the other two characteristics, it is still relevant.

Table 8. Feature importance table (Random Forest).

Characteristic	Importance
Syntactic structure	0.42
Behavior history	0.35
Demographics	0.23

For models such as logistic regression, the coefficients assigned to each characteristic can indicate their relevance in the prediction. Table 9 shows the coefficients associated with each characteristic in a logistic regression model and the respective values of the t statistic, which indicate the significance of each coefficient. “Syntactic Structure” has a positive coefficient of 2.45 and a t-statistic of 5.1. This suggests that as the salience of this feature increases, the model response is more likely to be positive, and its significance is high. On the other hand, “Behavioral history” presents a negative coefficient of -1.30 with a t-statistic of 4.5, which implies that an increase in this aspect could decrease the probability of a positive response. Still, it remains a significant predictor in the model. Finally, “Demographics” shows a positive coefficient of 0.75 and a t-statistic of 3.0, indicating a positive relationship of moderate magnitude with the model response and moderate significance compared to the other two characteristics.

Table 9. Table of coefficients (logistic regression).

Characteristic	Coefficient	T-Statistic
Syntactic structure	2.45	5.1
Behavior history	-1.30	4.5
Demographics	0.75	3.0

In permutation techniques, it is possible to infer its relevance by randomly changing the order of a feature and measuring how much it affects the performance of the model. Table 10 reflects the effects on model accuracy after randomly permuting the values of each feature. This analysis method helps us understand the relative importance of features by observing how much model performance deteriorates by altering the original information. “Syntactic structure” shows a 7% decrease in accuracy after permutation, suggesting that it is a crucial feature for the correct functioning of the model. Any perturbation in the syntactic structure leads to a significant degradation in the model’s predictive capacity.

Table 10. Table of change in post-permutation precision.

Characteristic	Importance
Syntactic structure	-7%
Behavior history	-5%
Demographics	-2%

“Behavior History” has a 5% reduction in precision, pointing out its relevance, although to a lesser degree than syntactic structure. This implies that the user’s past behavior is a valuable prediction indicator, and its alteration significantly affects the model’s performance. Finally, the permutation of “Demographics” results in a more modest 2% decrease in accuracy. Although this feature impacts predictions, it is less critical than the other two. This analysis underscores the importance of syntactic structure in the model, followed by behavioral history. At the same time, demographics play a more secondary role in the model’s overall accuracy.

3.4.3. Analysis Results

In an NLP-based chatbot, features that could be highlighted include the grammatically correct structure of the sentences, the identified keywords, and the length of the users’ messages. Also, contextual aspects include the time of day or the user’s history of previous interactions. In the analyses, it was found that keywords have a significant impact on the accuracy of the chatbot’s responses. Words that denote intent, such as “buy,” “reserve,” or “request,” are often decisive. The length of the message also proved to be a relevant factor. Longer, more descriptive messages tend to get more precise responses than short, ambiguous queries.

Table 11 summarizes the relative importance of various characteristics considered in the performance of an NLP-based chatbot. Keywords and grammatical structure are highly relevant in determining the effectiveness of the chatbot's responses. The length of the message and the history of previous interactions provide valuable context and are, therefore, of medium importance. Although the time of day can give specific contexts in certain applications, it generally has a minor importance in the overall accuracy of the chatbot.

Table 11. Table of relative importance of characteristics.

Characteristic	Importance
Syntactic structure	0.42
Behavior history	0.35
Demographics	0.23

3.5. Syntactic Analysis in Chatbots

A representative sample of various industries was chosen for syntactic analysis in chatbots to cover a spectrum of applications and broad-use contexts. For example, customer service assistants in retail were analyzed; these chatbots are programmed to handle product queries, inventory availability, and claims processing. Their ability to interpret questions with syntactic variations and provide appropriate answers in a dynamic e-commerce environment was analyzed.

Financial advisors in banks are designed to offer advice on banking products, investments, and financial services; these chatbots are evaluated on their ability to understand complex queries related to personal finances and provide accurate recommendations. Similarly, hospital health assistants were considered; these chatbots handle questions about medical appointments, treatment information, and general health advice. Their competency in processing specific medical terminology and responding coherently to complex health scenarios was assessed. In each case, key variables such as the frequency of interactions, the range of questions processed, and the tasks' complexity were considered. This diversity in applications and contexts provides a solid basis for evaluating the impact of syntax on the communicative effectiveness of chatbots.

First, questions of varying syntactic complexity were designed to evaluate the chatbots' responses, from simple queries to statements with intricate grammatical structures. These questions were presented to the chatbots, and their responses were recorded for analysis. For BERT, we focus on how the model processes and understands the context and intent behind the questions, evaluating its ability to correctly interpret syntax and implied meaning. With GPT, attention was directed to generating responses and observing the coherence, fluency, and relevance of the text produced in response to the questions posed. Each chatbot response was analyzed using a scoring framework that considered accuracy, relevance, and consistency, allowing for a quantitative and qualitative comparison between the different chatbots and the NLP tools used.

Table 12 presents ten questions of different syntactic complexity, classified as simple, moderate, and complex. Simple questions, such as asking the time or weather, received high ratings, indicating that the chatbots responded accurately and effectively. Moderate questions, such as requesting banking information or setting alarms, had a medium rating. Complex questions, which include conditional scenarios or requests for specific processes, showed a low rating, reflecting challenges in the chatbot's understanding and appropriate response.

Table 13 of response accuracy according to the NLP tool shows evaluation results for BERT and GPT in three categories of question complexity: simple, moderate, and complex. Each category groups questions of similar difficulty. For example, the two simple questions in the table above are grouped under 'Simple,' and the same goes for the 'Moderate' and 'Complex' categories. This explains why there are six results instead of ten. Accuracy percentages reflect each tool's effectiveness in answering questions in each category, with

BERT showing an advantage in complex queries and GPT excelling on simple and moderate questions. These results were obtained through a quantitative and qualitative analysis of the chatbots' responses to the questions.

Table 12. Chatbot response evaluation: classification by question complexity and accuracy.

Question Type	Question Example
High	"What time is it?"
High	"How is the weather today?"
Half	"I need information about my bank balance."
Half	"Can you set an alarm for 7 AM?"
Low	"If it rains, can I change my hotel reservation?"
Low	"What are the steps to request a refund?"
High	"Where is your nearest store located?"
Half	"I would like to compare prices of different products."
Low	"What options do I have if my flight is canceled?"
Low	"If my current plan doesn't include roaming, how can I add it?"

Table 13. Chatbot response accuracy comparison Table: evaluation with BERT and GPT.

NLP Tool	Complexity of the Question	Response Accuracy (%)
BERT	Simple	98
BERT	Moderate	88
BERT	Complex	72
GPT	Simple	96
GPT	Moderate	85
GPT	Complex	78

The results obtained reveal exciting patterns in the effectiveness of BERT and GPT. BERT showed higher accuracy on complex questions, likely due to its ability to better understand context and two-way sentence relationships. GPT, on the other hand, performed better on simple and moderate questions, which could be attributed to its focus on generating coherent and logically structured text. Problems encountered include GPT's difficulty in handling questions with complex implicit contexts and BERT's tendency to be less precise in answers to more direct questions. These findings suggest that while BERT is more suitable for situations that require a deep understanding of language, GPT is preferable for generating coherent responses in less complex situations.

4. Discussion

When analyzing the results obtained throughout this work, the profound influence of specific characteristics on the performance of a chatbot based on NLP is evident. The intersection between traditional machine learning metrics and the peculiarities of natural language processing has provided unique insight into how chatbots interpret and respond to user queries. When analyzing the results obtained throughout this work, the profound influence of specific characteristics on the performance of a chatbot based on NLP is evident. The intersection between traditional machine learning metrics and the peculiarities of natural language processing has provided unique insight into how chatbots interpret and respond to user queries. As highlighted in the study by Liu et al. [37], advanced techniques, such as visual semantic embedding regularization with contrastive learning, are critical to improving accuracy in natural language processing and human-computer interactions. This approach is particularly relevant to our discussion of the importance of syntax in chatbot training, as it highlights how the understanding and interpretation of requests can be optimized using sophisticated machine learning techniques. Recent studies indicate that chatbots trained with an emphasis on syntax show a more remarkable ability to distinguish between different types of requests, thus improving the relevance and accuracy of their responses. Case studies, especially in help desk environments, show

that chatbots with advanced syntactic interpretation skills can answer complex technical questions more accurately, reducing the need for human intervention and increasing customer satisfaction. This analysis highlights the importance of investing in developing natural language processing capabilities to optimize human–computer interactions.

Syntactic structure, for example, emerged as a dominant feature, reflecting the fundamental intuition that how a sentence is structured influences the interpretation of its meaning. These findings align with previous research emphasizing the importance of grammatical structure in NLP [38]. In reviewed work, it has been observed that models trained with a deep understanding of grammar tend to outperform those based solely on critical terms or direct matching approaches. On the other hand, the relevance of behavioral history highlights NLP's evolution towards more contextual systems [39]. A chatbot that can remember past interactions and adjust its responses accordingly is undoubtedly more effective. This trend has been reflected in contemporary research highlighting the need for more personalized and adaptive recommender systems. The findings of this study align with emerging trends in the field of NLP, which seek to develop more adaptable and personalized chatbot models. The ability to recalibrate and retrain the model based on the specific demands of the application domain is a feature that is becoming increasingly relevant in NLP research.

As for demographic data, its importance may initially seem minor compared to more direct linguistic characteristics. However, when considering the chatbot in a real-world scenario, these data can be essential to adjust the tone, style, and content of the response, especially in business applications where knowing the customer is crucial [40]. When contrasting our results with previous work, a clear trend is observed towards more holistic and contextual models in the field of NLP. Where rules-based or simple term-matching approaches once predominated, we now see a growing appreciation for context, history, and personalization [22,41].

The analysis techniques used also yielded promising results regarding the multi-faceted nature of NLP. While Decision Trees and logistic regression provided valuable insights into the importance and weight of features, the permutation technique highlighted the fragility and interdependence of these features. Changing a single feature, such as syntactic structure, could have a ripple effect on the accuracy and effectiveness of the chatbot [42,43]. Although our findings are robust within the defined context, they may not be generalized to all chatbot scenarios or applications. Additionally, the NLP field is constantly evolving, and what is true today may not be true tomorrow.

Although our findings are robust within the defined environment, they may not be generalizable to all chatbot scenarios or applications. This limitation arises from the diversity of human language and the varied ways dialogues can be structured in different applications. Our results are based on a specific set of data and model parameters. Although we have attempted to cover various scenarios, the findings may not apply universally to all chatbot applications. The diversity of human language and variability in dialogue structures represent a significant challenge to generalization. The effectiveness of our chatbot model depends closely on the quality and nature of the data set used. Chatbots developed for different domains or languages may require substantial adjustments, which was not fully addressed in our study [44].

We believe our findings are most directly applicable to chatbots designed for customer service. Accurate interpretation of queries and generating relevant responses are paramount in these contexts. Since our study primarily focused on Spanish and English data, the results are particularly relevant for chatbots operating in these languages. The models may need adaptations for languages with different syntactic structures.

To address these limitations and improve the external validity of our study, we propose that conducting additional studies in various domains and with different languages is crucial to evaluate and enhance the generalizability of the findings. Furthermore, it is essential to explore the adaptability of models to different linguistic and cultural structures, which is vital for creating universally effective chatbots.

Based on the reviewed works, it is crucial to mention that while our results largely align with the existing literature, there are discrepancies. Some research has found that characteristics we consider secondary, such as time of day, significantly impact specific contexts. This underscores the importance of considering domain specificity when designing and evaluating chatbots.

When approaching demographic data analysis in chatbot development, it is essential to consider ethical and privacy aspects carefully. The collection and use of demographic data must be carried out responsibly, ensuring that users' privacy and individual rights are protected. For this, it is crucial to obtain informed consent from users before collecting and analyzing their data. This includes transparency about data use and ensuring users know their rights.

Data must be anonymized, and appropriate security measures must be implemented to protect users' personal information. It is essential to ensure that the identity of individuals cannot be revealed through the data collected. Demographic data should be used only for the specified purposes and in a manner that does not promote bias or discrimination. The interpretation and application of these data must be carried out with cultural and social sensitivity.

5. Conclusions

A chatbot's ability to understand, interpret, and respond to user queries consistently and contextually is mainly attributable to advances in NLP. What was once a set of predefined or rule-based responses has evolved into more intuitive and adaptive systems that genuinely "understand" their users. Deep analysis of critical features such as syntactic structure, behavioral history, and demographics revealed their importance and how they interact with each other to influence chatbot responses. This level of granularity and understanding is essential to developing more sophisticated and effective chatbots in the future. Identifying and weighing these characteristics provides a clear roadmap for future research and development in chatbots.

Beyond individual characteristics, the analysis methods used in this study underline the importance of a holistic approach. It is not simply identifying which factors are essential but understanding how they relate to and influence each other. The permutation technique was revealing in this sense, highlighting the delicate interdependence of characteristics, and how modifying one can have repercussions on the overall performance of the chatbot.

Our work aligns with and extends several findings from previous research. Rather than viewing chatbots as isolated systems, we have placed them in the broader context of NLP, recognizing and addressing the inherent complexity of this field. Through rigorous analysis and evaluation, we have provided valuable insights for developers, researchers, and professionals in artificial intelligence and NLP. It is also essential to recognize the limitations and challenges that arose during the development of this study. Although we have obtained significant and robust results, there is always the risk that certain factors have not been considered or that rapid technological changes could alter the picture shortly. However, these limitations do not diminish the value of our work but rather underscore the need for continued and adaptive research in this field.

The analysis presented in this study not only highlights the specific contributions we have made but also sets a precedent for future work in this field. We have shown that, with the right approach and tools, it is possible to understand the complexities of NLP in the context of chatbots, bringing the scientific community one step closer to intuitive and humanized chatbots.

This study demonstrates the vital importance of syntactic elements in the development and efficiency of conversational agents. A detailed analysis of case studies has shown how an advanced understanding of syntax improves the accuracy and responsiveness of chatbots. Innovations in natural language processing, especially in syntactic interpretation, are crucial for achieving more natural interactions between humans and machines. This work

highlights the need to continue research focused on improving the syntactic capabilities of chatbots to improve user–chatbot interaction and open new paths in conversational artificial intelligence.

As for future work, as NLP continues to evolve, it is essential to maintain an adaptive approach and be willing to reevaluate and adjust our strategies. Furthermore, with the advent of cutting-edge artificial intelligence and deep learning systems, the potential for more advanced and contextual chatbots is immense. Exploring how these technologies can integrate and improve current systems will be crucial. Additionally, attention must be paid to ethics and privacy, ensuring that chatbots are practical and respectful of users' rights and sensitivities.

Author Contributions: Conceptualization, W.V.-C.; methodology, I.O.-G.; software, J.G.; validation, J.G.; formal analysis, I.O.-G.; investigation, R.O.A.; data curation, W.V.-C. and I.O.-G.; writing—original draft preparation, J.G. and R.O.A.; writing—review and editing, I.O.-G.; visualization, J.G.; supervision, W.V.-C. and R.O.A.; 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 (email to ivan.ortiz@udla.edu.ec).

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

References

- Ferrario, A.; Loi, M.; Viganò, E. In AI We Trust Incrementally: A Multi-Layer Model of Trust to Analyze Human-Artificial Intelligence Interactions. *Philos. Technol.* **2020**, *33*, 523–539. [CrossRef]
- Nee, J.; Smith, G.M.; Sheares, A.; Rustagi, I. Linguistic Justice as a Framework for Designing, Developing, and Managing Natural Language Processing Tools. *Big Data Soc.* **2022**, *9*, 20539517221090930. [CrossRef]
- Kang, Y.; Cai, Z.; Tan, C.W.; Huang, Q.; Liu, H. Natural Language Processing (NLP) in Management Research: A Literature Review. *J. Manag. Anal.* **2020**, *7*, 139–172. [CrossRef]
- Raharjana, I.K.; Siahaan, D.; Fatichah, C. User Stories and Natural Language Processing: A Systematic Literature Review. *IEEE Access* **2021**, *9*, 53811–53826. [CrossRef]
- Pham, K.T.; Nabizadeh, A.; Selek, S. Artificial Intelligence and Chatbots in Psychiatry. *Psychiatr. Q.* **2022**, *93*, 249–253. [CrossRef]
- Nawaz, N.; Saldeen, M.A. Artificial intelligence chatbots for library reference services. *J. Manag. Inf. Decis. Sci.* **2020**, *23*. Available online: <https://ssrn.com/abstract=3883917> (accessed on 10 July 2021).
- Skrebeca, J.; Kalniete, P.; Goldbergs, J.; Pitkevica, L.; Tihomirova, D.; Romanovs, A. Modern Development Trends of Chatbots Using Artificial Intelligence (AI). In Proceedings of the ITMS 2021-2021 62nd International Scientific Conference on Information Technology and Management Science of Riga Technical University, Proceedings, Riga, Latvia, 14–15 October 2021.
- Kulthe, S.; Tiwari, V.; Nirmal, M.; Chaudhari, B. Introspection of Natural Language Processing for Ai Chatbot. *Int. J. Technol. Res. Eng.* **2019**, *6*, 5178–5183.
- Costa, P.; Ribas, L. Ai Becomes Her: Discussing Gender and Artificial Intelligence. *Technoetic Arts* **2019**, *17*, 171–193. [CrossRef]
- Meshram, S.; Naik, N.; Megha, V.R.; More, T.; Kharche, S. College Enquiry Chatbot Using Rasa Framework. In Proceedings of the 2021 Asian Conference on Innovation in Technology, ASIANCON 2021, Pune, India, 27–29 August 2021.
- Lee, J.; Goodwin, R. Ontology Management for Large-Scale Enterprise Systems. *Electron. Commer. Res. Appl.* **2006**, *5*, 91–114. [CrossRef]
- Bona, J.P.; Prior, F.W.; Zozus, M.N.; Brochhausen, M. Enhancing Clinical Data and Clinical Research Data with Biomedical Ontologies-Insights from the Knowledge Representation Perspective. *Yearb. Med. Inf.* **2019**, *28*, 140–151. [CrossRef]
- Gruetzemacher, R. The Power of Natural Language Processing. Harvard Business Review Digital Article. 2022. Available online: <https://hbr.org/2022/04/the-power-of-natural-language-processing> (accessed on 6 August 2021).
- Laghrissi, F.E.; Douzi, S.; Douzi, K.; Hssina, B. Intrusion Detection Systems Using Long Short-Term Memory (LSTM). *J. Big Data* **2021**, *8*, 65. [CrossRef]
- Kratzert, F.; Klotz, D.; Brenner, C.; Schulz, K.; Herrnegger, M. Rainfall-Runoff Modelling Using Long Short-Term Memory (LSTM) Networks. *Hydrol. Earth Syst. Sci.* **2018**, *22*, 6005–6022. [CrossRef]
- Draaisma, L.R.; Wessel, M.J.; Moyne, M.; Morishita, T.; Hummel, F.C. Targeting the Frontoparietal Network Using Bifocal Transcranial Alternating Current Stimulation during a Motor Sequence Learning Task in Healthy Older Adults. *Brain Stimul.* **2022**, *15*, 968–979. [CrossRef]

17. Lim, M.H.; Zeng, A.; Ichter, B.; Bandari, M.; Coumans, E.; Tomlin, C.; Schaal, S.; Faust, A. Multi-Task Learning with Sequence-Conditioned Transporter Networks. In Proceedings of the IEEE International Conference on Robotics and Automation, Philadelphia, PA, USA, 23–27 May 2022.
18. Garcia-Silva, A.; Denaux, R.; Gomez-Perez, J.M. Learning Embeddings from Scientific Corpora Using Lexical, Grammatical and Semantic Information. In Proceedings of the CEUR Workshop Proceedings, Kryvyi Rih, Ukraine, 20 December 2019; Volume 2526.
19. Kreimeyer, K.; Foster, M.; Pandey, A.; Arya, N.; Halford, G.; Jones, S.F.; Forshee, R.; Walderhaug, M.; Botsis, T. Natural Language Processing Systems for Capturing and Standardizing Unstructured Clinical Information: A Systematic Review. *J. Biomed. Inform.* **2017**, *73*, 14–29. [[CrossRef](#)]
20. Casey, A.; Davidson, E.; Poon, M.; Dong, H.; Duma, D.; Grivas, A.; Grover, C.; Suárez-Paniagua, V.; Tobin, R.; Whiteley, W.; et al. A Systematic Review of Natural Language Processing Applied to Radiology Reports. *BMC Med. Inf. Decis. Mak.* **2021**, *21*, 179. [[CrossRef](#)]
21. Rezaei, N.; Wolff, P.; Price, B.H. Natural Language Processing in Psychiatry: The Promises and Perils of a Transformative Approach. *Br. J. Psychiatry* **2022**, *220*, 251–253. [[CrossRef](#)]
22. Álvarez-Carmona, M.; Aranda, R.; Rodríguez-Gonzalez, A.Y.; Fajardo-Delgado, D.; Sánchez, M.G.; Pérez-Espinosa, H.; Martínez-Miranda, J.; Guerrero-Rodríguez, R.; Bustio-Martínez, L.; Díaz-Pacheco, A. Natural Language Processing Applied to Tourism Research: A Systematic Review and Future Research Directions. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 10125–10144. [[CrossRef](#)]
23. Zhou, B.; Yang, G.; Shi, Z.; Ma, S. Natural Language Processing for Smart Healthcare. *IEEE Rev. Biomed. Eng.* **2022**, *17*, 4–8. [[CrossRef](#)]
24. Zhang, W.E.; Sheng, Q.Z.; Alhazmi, A.; Li, C. Adversarial Attacks on Deep-Learning Models in Natural Language Processing. *ACM Trans. Intell. Syst. Technol.* **2020**, *11*, 1–41. [[CrossRef](#)]
25. Garcia-Teruel, R.M.; Simón-Moreno, H. The Digital Tokenization of Property Rights. A Comparative Perspective. *Comput. Law Secur. Rev.* **2021**, *41*, 105543. [[CrossRef](#)]
26. Omar, M.; Choi, S.; Nyang, D.; Mohaisen, D. Robust Natural Language Processing: Recent Advances, Challenges, and Future Directions. *IEEE Access* **2022**, *10*, 86038–86056. [[CrossRef](#)]
27. Nguyen, H.M.; Miyazaki, T.; Sugaya, Y.; Omachi, S. Multiple Visual-Semantic Embedding for Video Retrieval from Query Sentence. *Appl. Sci.* **2021**, *11*, 3214. [[CrossRef](#)]
28. Yeh, M.C.; Li, Y.N. Multilabel Deep Visual-Semantic Embedding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 1530–1536. [[CrossRef](#)]
29. Gong, Y.; Cosma, G. Improving Visual-Semantic Embeddings by Learning Semantically-Enhanced Hard Negatives for Cross-Modal Information Retrieval. *Pattern Recognit.* **2023**, *137*, 109272. [[CrossRef](#)]
30. Zhu, L.; Shen, J.; Gartner, G.; Hou, Y. Personalized Landmark Sequence Recommendation Method Using LSTM-Based Network for Navigating in Large Hospitals. *Abstr. ICA* **2021**, *3*, 1–2. [[CrossRef](#)]
31. Aguayo, R.; Lizarraga, C.; Quiñonez, Y. Evaluation of Academic Performance in Virtual Environments Using the Nlp Model. *RISTI-Rev. Iber. De Sist. E Tecnol. De Inf.* **2021**, *2021*, 34–49. [[CrossRef](#)]
32. Shadiev, R.; Wu, T.T.; Huang, Y.M. Using Image-to-Text Recognition Technology to Facilitate Vocabulary Acquisition in Authentic Contexts. *ReCALL* **2020**, *32*, 195–212. [[CrossRef](#)]
33. Niehorster, D.C.; Li, L.; Lappe, M. The Accuracy and Precision of Position and Orientation Tracking in the HTC Vive Virtual Reality System for Scientific Research. *Iperception* **2017**, *8*, 2041669517708205. [[CrossRef](#)]
34. DeVries, Z.; Locke, E.; Hoda, M.; Moravek, D.; Phan, K.; Stratton, A.; Kingwell, S.; Wai, E.K.; Phan, P. Using a National Surgical Database to Predict Complications Following Posterior Lumbar Surgery and Comparing the Area under the Curve and F1-Score for the Assessment of Prognostic Capability. *Spine J.* **2021**, *21*, 1135–1142. [[CrossRef](#)]
35. Carrón, J.; Campos-Roca, Y.; Madruga, M.; Pérez, C.J. A Mobile-Assisted Voice Condition Analysis System for Parkinson’s Disease: Assessment of Usability Conditions. *Biomed. Eng. Online* **2021**, *20*, 114. [[CrossRef](#)]
36. Zhang, P.; Shi, X.; Khan, S.U.; Ferreira, B.; Portela, B.; Oliveira, T.; Borges, G.; Domingos, H.; Leitão, J.; Mohottige, I.P.; et al. IEEE Draft Standard for Spectrum Characterization and Occupancy Sensing. *IEEE Access* **2019**, *9*.
37. Liu, Y.; Liu, H.; Wang, H.; Liu, M. Regularizing Visual Semantic Embedding With Contrastive Learning for Image-Text Matching. *IEEE Signal Process. Lett.* **2022**, *29*, 1332–1336. [[CrossRef](#)]
38. Le Glaz, A.; Haralambous, Y.; Kim-Dufor, D.H.; Lenca, P.; Billot, R.; Ryan, T.C.; Marsh, J.; DeVylder, J.; Walter, M.; Berrouguet, S.; et al. Machine Learning and Natural Language Processing in Mental Health: Systematic Review. *J. Med. Internet Res.* **2021**, *23*, e15708. [[CrossRef](#)]
39. Juhn, Y.; Liu, H. Artificial Intelligence Approaches Using Natural Language Processing to Advance EHR-Based Clinical Research. *J. Allergy Clin. Immunol.* **2020**, *145*, 463–469. [[CrossRef](#)]
40. Shaik, T.; Tao, X.; Li, Y.; Dann, C.; McDonald, J.; Redmond, P.; Galligan, L. A Review of the Trends and Challenges in Adopting Natural Language Processing Methods for Education Feedback Analysis. *IEEE Access* **2022**, *10*, 56720–56739. [[CrossRef](#)]
41. Cai, M. Natural Language Processing for Urban Research: A Systematic Review. *Heliyon* **2021**, *7*, e06322. [[CrossRef](#)]
42. Shoenbill, K.; Song, Y.; Gress, L.; Johnson, H.; Smith, M.; Mendonca, E.A. Natural Language Processing of Lifestyle Modification Documentation. *Health Inform. J.* **2020**, *26*, 388–405. [[CrossRef](#)] [[PubMed](#)]

43. Cîmpeanu, I.-A.; Dragomir, D.-A.; Zota, R.D. Banking Chatbots: How Artificial Intelligence Helps the Banks. *Proc. Int. Conf. Bus. Excell.* **2023**, *17*, 1716–1727. [[CrossRef](#)]
44. Ucak, U.V.; Ashyrmamatov, I.; Lee, J. Improving the Quality of Chemical Language Model Outcomes with Atom-in-SMILES Tokenization. *J. Cheminform.* **2023**, *15*, 55. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.