# Exploring the Expressive Power of Large Language Models in Neuro-Fuzzy System Explainability: A Study on EEG-Based Seizure Detection

Gabriella Casalino*, Giovanna Castellano, Daniele Margherita, Alberto Gaetano Valerio, Gennaro Vessio and Gianluca Zaza

*Department of Computer Science, University of Bari Aldo Moro, Bari, Italy*

**Abstract**

In this work, we focus on integrating LLMs into a neuro-symbolic framework to enhance the quality of explanations associated with IF-THEN rules generated by neuro-fuzzy inference systems. To address the challenge posed by the lack of a reference ground truth in explanation tasks, we propose a quantitative evaluation based on linguistic and semantic quality metrics, aiming to assess the clarity, coherence, and relevance of the generated text. We systematically compare a selection of LLMs varying in size and architectural family, and investigate the impact of different prompting strategies—including zero-shot, persona-based, and fact-checking approaches—on the resulting explanations. The proposed framework is applied to a real-world case study on EEG-based seizure detection, illustrating its potential in high-stakes medical contexts where transparency and reliability are critical. The findings show that quantitative metrics alone are insufficient to capture the true quality of explanations, highlighting the critical role of both model selection and prompt design in generating effective, trustworthy, and human-aligned explanations.

**Keywords**

Explainable Artificial Intelligence, Large Language Models, Neuro-Fuzzy Systems, EEG-based Seizure Detection

## 1. Introduction

In recent years, Large Language Models (LLMs) have emerged as powerful AI tools trained on massive text corpora, enabling them to encode and generalize linguistic knowledge. Their broad applicability stems from this rich pretraining, allowing effective use across diverse domains, such as healthcare [1, 2], education [3], software engineering [4], and human capital management [5], among others, often without the need for task-specific fine-tuning. LLMs are increasingly important in the medical domain, where they support a variety of tasks such as clinical decision-making, medical documentation, patient triage, and question answering, thanks to their ability to process and generate complex biomedical language [6, 7].

However, in the medical domain, the adoption of AI necessitates transparent and interpretable explanations to foster trust, ensure accountability, and enable safe integration into clinical practice. Clinicians must be able to understand and justify AI-driven decisions, particularly in high-stakes scenarios [8]. Explainability not only aids domain experts in validating model outputs but also promotes acceptance and effective human–AI collaboration [9, 10].

Although LLMs are inherently black-box models, they are increasingly being adopted in the state of the art for generating explanations. Current research explores the use of in-context learning to inject both domain-specific and explainability-related knowledge into LLMs, enabling them to generate responses that combine user-friendly narratives for non-experts with technical insights for specialists

*EXPLIMED 2025 - Second Workshop on Explainable Artificial Intelligence for the medical domain - 25-30 October 2025, Bologna, Italy*

*Corresponding author.

gabriella.casalino@uniba.it (G. Casalino); giovanna.castellano@uniba.it (G. Castellano); d.margherita@studenti.uniba.it (D. Margherita); a.valerio31@phd.uniba.it (A. G. Valerio); gennaro.vessio@uniba.it (G. Vessio); gianluca.zaza@uniba.it (G. Zaza)

0000-0003-0713-2260 (G. Casalino); 0000-0002-6489-8628 (G. Castellano); 0009-0008-6101-965X (A. G. Valerio); 0000-0002-0883-2691 (G. Vessio); 0000-0003-3272-9739 (G. Zaza)

[11]. In many frameworks, LLMs are integrated with traditional machine learning models to transform outputs into textual explanations that are closer to the user's perspective [12, 13]. This integration is sometimes further enhanced through external resources, such as knowledge graphs, which provide additional contextualization and contribute to more informative and trustworthy explanations [14].

In contrast to black-box machine learning models, fuzzy logic has long proven valuable for explainability, as it is inherently transparent and enables the representation of human perceptions through linguistic terms that map numerical values onto the imprecise concepts characteristic of human reasoning [15]. Neuro-fuzzy inference systems further extend this capability by automatically extracting fuzzy sets from data and generating IF–THEN rules that replicate human-like reasoning [16]. This paradigm supports a natural transition from rigid, crisp representations to more flexible models that inherently accommodate the uncertainty and vagueness embedded in human language and thought.

In a recent study [17] , we explored the use of LLMs to generate context-aware textual explanations and proposed a novel workflow that integrates neuro-fuzzy systems with LLMs. This integration improved the resulting explanations by making them more transparent, structured, and coherent. Importantly, the workflow was designed within a human-in-the-loop, human-centered explainability framework, positioning domain experts at the core of the explanation process [18]. This hybrid approach not only bridges symbolic and sub-symbolic reasoning but also promotes expert engagement, interpretability, and trust, thereby aligning system outputs with human cognitive and decision-making processes [17]. Despite several applications combining fuzzy logic with LLMs [19, 20, 21], this was among the first attempts to explicitly leverage LLMs within fuzzy logic–based explainability [22]. The workflow was evaluated through a case study on EEG-based seizure classification, a high-stakes problem in epilepsy care where transparent, expert-aligned explanations can reduce misdiagnosis and guide timely treatment.

In this work, we extend our previous study by comparing LLMs from different families and sizes. We quantitatively evaluate the generated explanations using metrics that assess textual quality without relying on a ground truth, which is typically unavailable in this type of task. Furthermore, we investigate the effectiveness of different prompt patterns, including zero-shot prompting, persona-based prompting, and fact-checking–oriented prompting. Particularly, this work aims to address the following research questions, focusing on the effectiveness and quality of LLM-generated explanations in the context of neuro-symbolic systems applied to the same predictive task:

RQ1 Which LLMs generate the most effective explanations, based on linguistic quality criteria such as fluency, coherence, and well-formedness?

RQ2 Is there a prompt pattern, among zero-shot, persona, persona + fact-checking, that yields higher-quality explanations in terms of linguistic clarity, coherence, and contextual relevance?

The rest of the paper is organized as follows: Section 2 presents the experimental setup, Section 3 discusses the results, and Section 4 outlines the conclusions and directions for future work.

## 2. Matherials and Methods

To address the research questions outlined above, we conducted a case study on EEG-based epileptic seizure classification. Building on the pipeline introduced in [17], we first extracted a fuzzy rule base—an interpretable set of IF–THEN rules with fuzzy predicates capturing graded relationships—from EEG signals and subsequently applied a range of LLMs, combined with different prompting strategies, to assess and compare their ability to generate meaningful textual explanations. The quality of the generated explanations was evaluated using quantitative linguistic and semantic metrics. The overall workflow is summarized in Figure 1 and described in detail in the following sections. The proposed pipeline consists of three main phases:

1. Computational phase: This phase involves the training and evaluation of an Adaptive Neuro-Fuzzy Inference System (ANFIS) [23] functioning as a neuro-fuzzy predictive model, accompanied by the extraction of fuzzy rules and the use of unsupervised clustering to select representative data samples for each class.

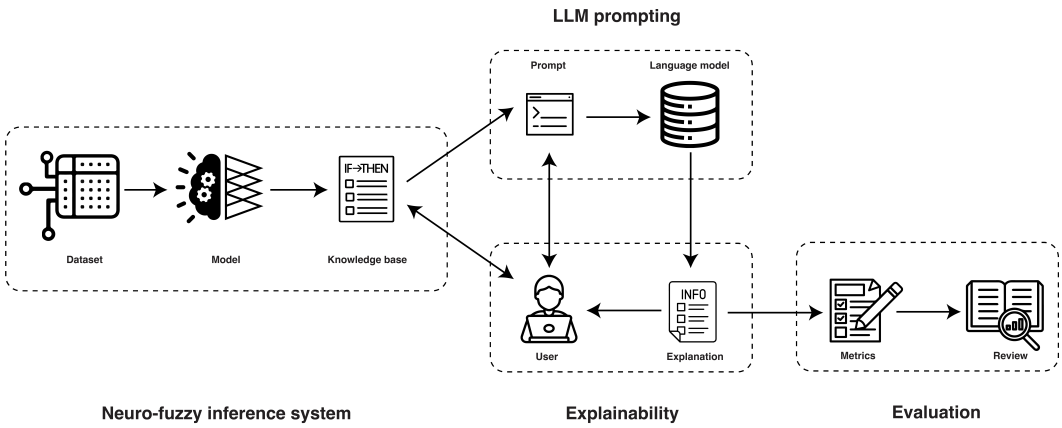



**Figure 1:** Proposed pipeline.

2. Explanatory phase: This phase focuses on transforming symbolic information into interpretable textual descriptions. The extracted fuzzy rules are reformulated by replacing generic labels with semantically meaningful names and then used to construct targeted prompts for LLMs, generating human-centered and understandable explanations.
3. Evaluation phase: In this phase, the explanations generated by the LLMs are analyzed and compared using linguistic quality metrics.

Unlike our previous work, this study does not focus extensively on the user-centered aspects of the pipeline, even if these remain relevant and will be addressed in future developments involving domain expert engagement. Instead, we concentrate on the evaluation module and the construction of the fuzzy rule base, which are central to the current investigation.

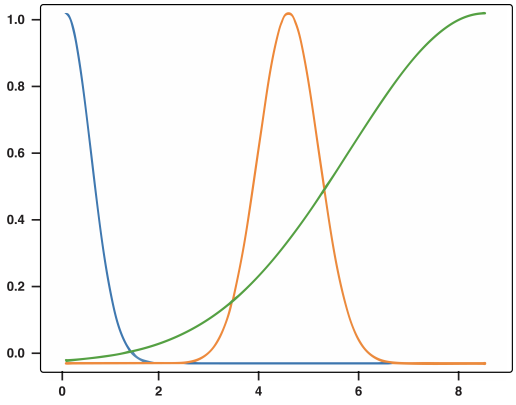
## 2.1. Data

Epilepsy is one of the most prevalent neurological disorders worldwide, affecting over 50 million people, with approximately one-third of patients continuing to experience frequent seizures despite treatment with multiple antiepileptic drugs [24]. EEG, a non-invasive technique for recording brain activity, is widely used in clinical settings for the diagnosis of epilepsy. However, EEG signals are high-dimensional, non-stationary, and often noisy, making their interpretation challenging and subjective. To address these issues, automated systems based on AI have gained prominence, offering promising tools for accurate and efficient seizure detection [25].

We used the *Epilepsy2* dataset [26] to assess the applicability of the proposed pipeline. This dataset includes single-channel EEG recordings from 500 individuals, each with a duration of 23.6 seconds. The recordings were segmented into 11,500 one-second intervals , sampled at 178 Hz, and randomly shuffled. The classification was framed as a binary task: class 0 represented epileptic seizure events, and class 1 denoted non-epileptic segments.

The EEG signals underwent preprocessing through Fourier transformation, which enabled the decomposition of each signal into five principal frequency bands. These bands are known to reflect different neurological and cognitive states, and their analysis aids in highlighting patterns associated with epileptic activity:

- Alpha (8–12 Hz): Typically observed during states of calm wakefulness with closed eyes, linked to a relaxed mental state.
- Beta (12–30 Hz): Characteristic of focused attention and active mental engagement, often present during movement and cognitive tasks.
- Gamma (30–50 Hz): Involved in complex cognitive processing, including sensory perception and higher-order brain functions.




(a)

Rule 7: IF DELTA is LOW and THETA is LOW and ALPHA is LOW and BETA is HIGH and GAMMA is LOW THEN EPILEPTIC

Rule 239: IF DELTA is HIGH and THETA is HIGH and ALPHA is HIGH and BETA is MEDIUM and GAMMA is MEDIUM THEN EPILEPTIC

Rule 240: IF DELTA is HIGH and THETA is HIGH and ALPHA is HIGH and BETA is MEDIUM and GAMMA is HIGH THEN EPILEPTIC

Rule 241: IF DELTA is HIGH and THETA is HIGH and ALPHA is HIGH and BETA is HIGH and GAMMA is LOW THEN NON-EPILEPTIC

Rule 242: IF DELTA is HIGH and THETA is HIGH and ALPHA is HIGH and BETA is HIGH and GAMMA is MEDIUM THEN NON-EPILEPTIC

(b)

**Figure 2:** Example of fuzzy sets *Low* (blue), *Medium* (orange), and *High* (green) for the fuzzy variable *Theta*; the x-axis represents *Theta* values and the y-axis the membership degree (a), alongside an excerpt of the ANFIS-generated fuzzy rule base (b).

- Delta (0.5–4 Hz): Dominant during deep sleep, indicating phases of physical and neural restoration.
- Theta (4–8 Hz): Associated with drowsiness, light sleep, and introspective activities such as meditation or imagination.

Since the original training set was imbalanced, with 1,630 non-epileptic and 6,432 epileptic instances, resulting in two balanced classes of 6,432 instances each. This step was crucial not only to ensure fair model training but also to prevent the majority class from dominating the generated explanations, which could lead to misleading or biased interpretations.

## 2.2. Computational phase

This phase represents the core of the modeling process, where raw EEG data are preprocessed and balanced to train an Adaptive Neuro-Fuzzy Inference System, which is employed to perform the classification task.

ANFIS is a four-layer feed-forward architecture that embeds fuzzy logic into a trainable neural network. ANFIS models the input space using Gaussian membership functions and learns a set of zero-order Takagi-Sugeno (TS) rules, where each rule associates a combination of fuzzy conditions with a constant output value. Formally, the TS rules can be expressed as:

$$\text{IF } (x_1 \text{ is } A_{k1}) \text{ AND } \ldots \text{ AND } (x_n \text{ is } A_{kn}) \text{ THEN } (y \text{ is } b_k),$$

where $k = 1, \ldots, K$ represents the rule index, $K$ is the total number of rules, $n$ is the number of fuzzy variables, $A_{ki}$ denotes a fuzzy set for the input variable $x_i$ in the $k$-th rule, and $b_k$ is the fuzzy singleton defining the output class.

The system computes the degree of membership, the rule activation (using a product t-norm), and the final output through a weighted sum of the rule consequents. Parameters are optimized via backpropagation, allowing the network to adapt while maintaining a transparent, rule-based structure.

In the context of the seizure classification, for each of the five EEG features describing the data (Alpha, Beta, Gamma, Delta, Theta), ANFIS automatically generated three Gaussian fuzzy sets from the data, corresponding to the linguistic terms Low, Medium, and High, as shown in Figure 2b. The consequent of each rule is a numerical singleton representing the class label (Epileptic or Non-Epileptic). Figure 2b shows an excerpt of the fuzzy rule base, generated after training the neuro-fuzzy network.

The total number of rules in ANFIS depends on the number of fuzzy variables and the number of fuzzy terms used to describe them. In this case, the rule base consists of 243 rules, since the total is obtained by raising the number of linguistic terms to the power of the number of fuzzy variables.

After training the model and generating the fuzzy rule base, the activation degree of each rule is computed for every test instance. These raw values are then normalized using an $\ell_1$ (unit-sum) normalization [27], ensuring that the activations for each instance sum to 1:

$$\tilde{\mu}r(x) = \frac{\mu_r(x)}{\sum j = 1^R \mu_j(x)}, \tag{1}$$

where $\mu_r(x)$ denotes the raw activation of rule $r$ for instance $x$, and $R$ is the total number of fuzzy rules ($R = 243$). For each instance, the rule with the highest normalized activation is identified as the `DominantRule`, following a winner-takes-all strategy [28]:

$$\text{DominantRule}(x) = \arg\max_r \tilde{\mu}_r(x). \tag{2}$$

We identified three examples from the epileptic class and three from the non-epileptic class using k-means clustering, selecting the instances closest to the cluster centroids. This procedure yields a representative and compact subset of instances that captures the central regions of each cluster. These instances are then used in the subsequent explanation phase based on fuzzy rules.

## 2.3. Explanatory phase

After analyzing the data and identifying the most active fuzzy rules in key regions of the decision space, we initiated the explanatory phase. The goal of this phase is to transform the symbolic information generated by the system—such as fuzzy rules—into interpretable textual descriptions expressed in natural language.

This phase consists of two complementary steps:

1. Fuzzy rule explication, i.e., translating raw rules into a semantically meaningful and readable format, accessible also to clinicians and non-technical users;
2. Automatic explanation generation, carried out through Large Language Models instructed to provide coherent, human-centered, and context-aware descriptions.

In the first step, the fuzzy rules generated by the ANFIS model are initially expressed in a generic form, with references to numerical variables (`x0`, `x1`, ..., `x4`) and unlabeled membership functions (`mf0`, `mf1`, `mf2`). To improve interpretability and clarify the semantic meaning of each rule, we introduced an explicitation phase. In this step, symbolic placeholders were reformulated by replacing the numerical variables with their corresponding EEG frequency bands (`x0` = Delta, `x1` = Theta, `x2` = Alpha, `x3` = Beta, `x4` = Gamma). Similarly, the membership functions were mapped to meaningful linguistic labels, with `mf0` corresponding to Low, `mf1` to Medium, and `mf2` to High. Finally, the output labels were made explicit by associating the vector `[1.0, 0.0]` with the class Non-Epileptic and `[0.0, 1.0]` with the class Epileptic.

For the second step, dedicated to the automatic generation of explanations, we employed different LLMs and tested three prompt patterns to compare their effectiveness in producing meaningful explanations.

### 2.3.1. Prompt engineering

Prompt engineering plays a crucial role in guiding language models to produce accurate, coherent, and contextually meaningful explanations. It involves the deliberate design of input prompts to align model outputs with user intent and task-specific requirements [29]. In this study, inspired by the catalog of patterns proposed by White et al. [30], we adopted three prompting strategies, zero-shot prompting, the persona pattern, and fact-checking, to evaluate their effectiveness in explanation generation.

The zero-shot prompt involves presenting the language model with a task-specific instruction without providing any in-prompt examples. As illustrated in Figure 3a, the prompt includes a brief clinical context (e.g., the patient's condition), a statement indicating that the result was produced using a fuzzy

inference system, and, most importantly, the fuzzy rule that was most activated for the given instance. This structure enables the LLM to grasp the key concepts expressed by the fuzzy rule that contribute to the decision. The model must rely exclusively on its pre-trained knowledge to interpret the prompt and generate a coherent and informative explanation.

This approach enables us to assess the model's generalization capabilities and its ability to follow natural language instructions without requiring fine-tuning or in-context examples. The expected output is a descriptive, user-friendly explanation of the selected fuzzy rule, enriched with relevant medical knowledge—especially concerning the condition and symptoms described in the rule—thus enhancing the system's interpretability for end-users.

In this work, we also investigate the effectiveness of the persona pattern (Figure 3b), a prompt engineering strategy in which the language model is instructed to adopt a specific role, identity, or perspective, referred to as a "persona", to influence the tone, style, and content of its responses. By embedding a persona within the prompt, users can control the level of formality and tone, simulate domain-specific expertise, and enhance coherence and consistency, especially in multi-turn interactions. This pattern is particularly valuable in tasks involving explainability, such as the medical domain. In our setting, we instructed the language model to assume the role of a physician and generate explanations in a communication style that is more familiar and accessible to end users (e.g., healthcare professionals or patients). Our objective is to assess whether adopting this pattern leads to more effective and user-centered explanations compared to the zero-shot baseline.

Finally, we experimented with a fact-checking pattern (Figure 3c) designed to enhance the transparency of the generated output. Since language models are known to produce plausible but factually unsupported statements, this strategy extends the persona pattern, where the model adopts the role of a physician, by explicitly requiring a concise list of key scientific or clinical facts at the end of each explanation. This "fact list" allows users to verify the reliability of the information against external sources independently. Although it does not eliminate the risk of errors, this strategy introduces an additional layer of verifiability and accountability in the use of LLMs in the medical domain.

### 2.3.2. Large Language Models

We conducted a comparative analysis of 14 LLMs, selected to represent a wide range of model families, architectural designs, and training strategies. We aimed to assess both intra-family variations—by comparing models with different architectures within the same family—and inter-family differences across distinct LLM frameworks. All selected models are open-source and were chosen based on their compatibility with local execution, given the computational constraints of running experiments on a standard laptop equipped with an 11th-generation Intel Core i7 CPU, 16 GB RAM, an NVIDIA RTX 3050 GPU with 4 GB of VRAM, and a 2 TB SSD. Below is a brief overview of each model considered in the study:

- LLaMA 3 (8B and 70B, 8192 context) [31]: Advanced models from Meta's LLaMA family, optimized for extended context handling and strong performance in generative and reasoning tasks.
- LLaMA 3.1-8B-Instant [31]: A latency-optimized variant designed for rapid response generation without significant loss in output quality.
- LLaMA 3.3-70B-Versatile [31]: A robust, general-purpose model aimed at multitask adaptability and long-form generation coherence.
- Allam-2-7B [32]: A lightweight yet expressive model, suitable for low-latency applications and edge computing scenarios.
- DeepSeek R1 Distill LLaMA-70B [33]: A distilled version of the LLaMA-70B designed to reduce inference cost while maintaining accuracy, particularly in summarization and reasoning.
- LLaMA-4 Maverick 17B (128e) [34]: An extended-embedding model fine-tuned for complex tasks, with emphasis on explainability and controllable outputs.
- LLaMA-4 Scout 17B (16e) [35]: Tailored for efficient interpretation and classification tasks, with optimizations for linguistic clarity.

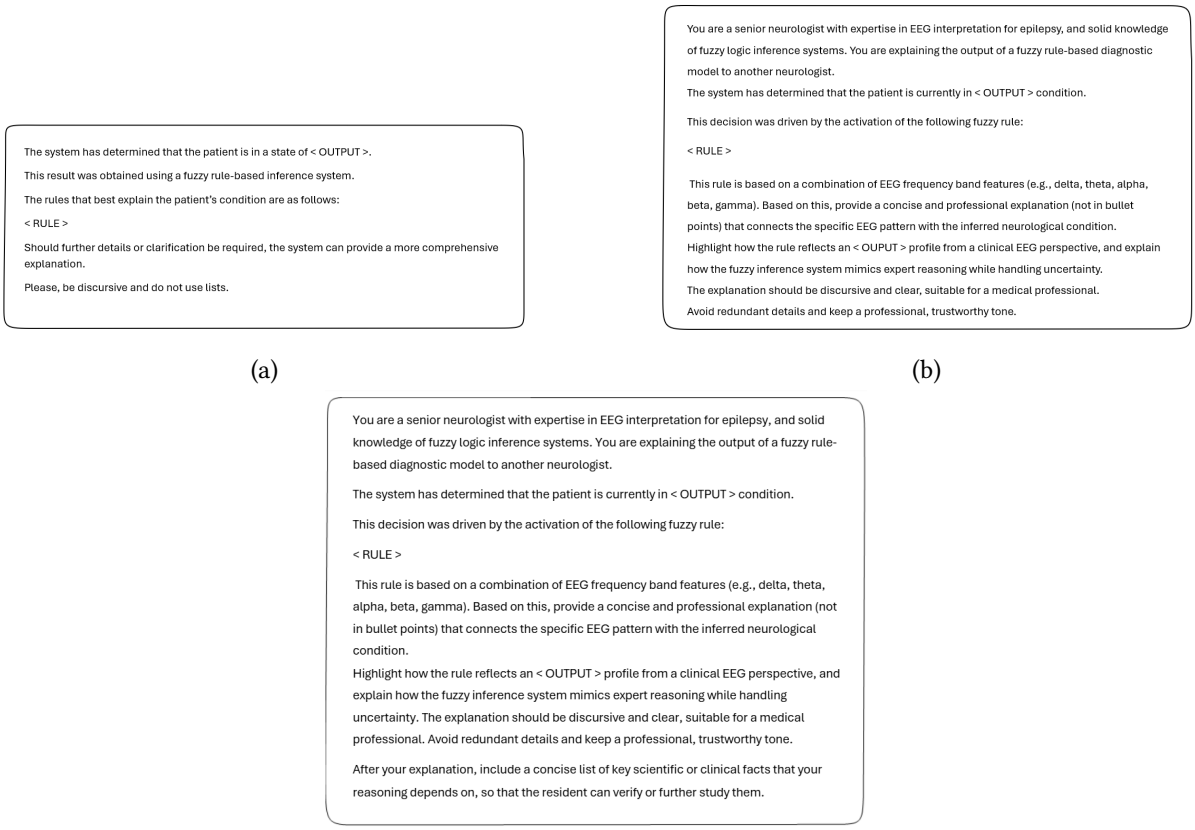



**Figure 3:** Templates of the prompts used in this study: (a) zero-shot pattern, and (b) persona and (c) fact-checking patterns, adopted to guide LLMs in the generation of textual explanations.

- Qwen-QWQ-32B [36]: A multilingual model by Alibaba, demonstrating robustness in noisy settings and solid performance on semantic tasks.
- Compound-Beta [37] and Compound-Beta-Mini [38]: Experimental models focused on merging generative and logic-based components to improve transparency and interpretability.
- Google Flan-T5 (Small, Base, Large) [39]: Instruction-tuned encoder-decoder models based on T5, known for their efficiency in structured tasks such as classification, translation, and explanation.

## 2.4. Evaluation phase

To quantitatively assess the quality of the generated explanations, we defined a multifaceted evaluation framework including linguistic, semantic, and computational aspects. The dimensions considered were lexical diversity, readability, coherence, information coverage, and generation time. All metrics were computed using the Spacy library [40], complemented with custom implementations for semantic similarity and coverage analysis.

Lexical diversity was assessed using two widely adopted indicators. The first is the *Type-Token Ratio (TTR)*, which captures the richness of vocabulary as the proportion of unique words (types) to the total number of words (tokens) in a text:

$$\text{TTR} = \frac{|\text{types}|}{|\text{tokens}|}. \tag{3}$$

A higher TTR reflects a more varied and informative lexical composition. To account for text length sensitivity, we also employed the *Maas Index*, a logarithmic transformation of TTR that adjusts for increasing token counts:

$$\text{Maas} = \frac{\log(|\text{tokens}|) - \log(|\text{types}|)}{(\log(|\text{tokens}|))^2}. \tag{4}$$

Lower values of the Maas Index indicate greater lexical diversity, offering a more robust assessment across explanations of varying lengths.

Readability was quantified through the *Flesch Reading Ease Score (FRES)*, which evaluates the comprehensibility of a text based on average sentence length and syllable density. The formula is defined as:

$$\text{FRES} = 206.835 - (1.015 \times \text{ASL}) - (84.6 \times \text{ASW}), \tag{5}$$

where ASL is the average sentence length and ASW is the average number of syllables per word. Higher FRES values denote more readable texts, a key factor in user-centered medical explanation.

To evaluate coherence, we adopted a metric based on *cosine similarity between adjacent sentence embeddings*, capturing the logical flow of information across the explanation. Given a text composed of $N$ sentences with embeddings $S_1, \ldots, S_N$, the coherence score is computed as:

$$\text{CohS} = \frac{1}{N-1} \sum_{i=1}^{N-1} \cos(S_i, S_{i+1}), \tag{6}$$

where $\cos(S_i, S_{i+1})$ denotes the cosine similarity between sentence $i$ and sentence $i+1$. Higher values suggest smoother transitions and a more logically consistent narrative.

We also evaluated the coverage of information through two complementary metrics. The first is the *Embedding-based Coverage Score (ECS)*, which measures the alignment between the generated explanation and the prompt by averaging the cosine similarity between the explanation embedding and each reference sentence embedding:

$$\text{ECS} = \frac{1}{N-1} \sum_{i=1}^{N-1} \cos(S_{\text{text}}, S_{\text{ref}_i}), \tag{7}$$

where $S_{\text{text}}$ is the embedding of the LLM-generated output, and $S_{\text{ref}_i}$ is the embedding of the $i$-th sentence in the prompt. The second is the *Token-based Coverage Score (TCS)*, which quantifies lexical overlap between the prompt and the generated explanation. After removing punctuation and stop words, the TCS is computed as:

$$\text{TCS} = \frac{|T_{\text{text}} \cap T_{\text{ref}}|}{|T_{\text{text}} \cup T_{\text{ref}}|}, \tag{8}$$

where $T_{\text{text}}$ and $T_{\text{ref}}$ are the sets of tokens from the output and the prompt, respectively. Higher TCS values indicate that more relevant information from the prompt is preserved in the generated explanation.

Finally, we recorded the generation time, defined as the number of seconds elapsed from prompt submission to completion of the model's output. This metric reflects computational efficiency, which is especially relevant for real-time or interactive applications.

## 3. Results and Discussion

In this section, we present the quantitative results obtained from the explanations generated by the 14 considered LLMs, using three prompting patterns across six examples (three from the positive class and three from the negative class). The evaluation relies on the measures previously described, with average values reported to provide a general overview of the performance of each model.

To facilitate the interpretation of the results, we summarized the tabular data into radar charts, highlighting the five analyzed dimensions: lexical diversity, readability, coherence, information coverage, and generation time. Lexical diversity was computed as:

$$\text{Diversity} = \frac{\text{TTR} + (1 - \text{MAAS})}{2}, \tag{9}$$

while coverage was defined as:

$$\text{Coverage} = \frac{\text{ECS} + \text{TCS}}{2}. \tag{10}$$

**Table 1**
Quantitative evaluation of explanations generated with the zero-shot prompt pattern across different LLMs.

| Model | TTR | Maas | FRES | CohS | ECS | TCS | Time |
|---|---|---|---|---|---|---|---|
| llama3-8b-8192 | 0.53 | 0.03 | 30.38 | 0.52 | 0.43 | 0.27 | 3.50 |
| llama-3.1-8b-instant | 0.47 | 0.03 | 26.15 | 0.50 | 0.40 | 0.25 | 1.19 |
| llama-3.3-70b-versatile | 0.49 | 0.03 | 22.44 | 0.48 | 0.42 | 0.26 | 2.03 |
| llama3-70b-8192 | 0.68 | 0.02 | 30.81 | 0.45 | 0.40 | 0.32 | 1.34 |
| allam-2-7b | 0.63 | 0.02 | 30.57 | 0.33 | 0.39 | 0.22 | 0.85 |
| deepseek-r1-distill-llama-70b | 0.50 | 0.02 | 41.50 | 0.35 | 0.41 | 0.10 | 4.74 |
| meta-llama/llama-4-maverick-17b-128e-instruct | 0.60 | 0.02 | 22.38 | 0.45 | 0.46 | 0.16 | 2.64 |
| meta-llama/llama-4-scout-17b-16e-instruct | 0.57 | 0.02 | 28.30 | 0.45 | 0.43 | 0.16 | 1.73 |
| qwen-qwq-32b | 0.58 | 0.01 | 33.47 | 0.38 | 0.43 | 0.08 | 3.02 |
| compound-beta | 0.52 | 0.02 | 23.88 | 0.49 | 0.46 | 0.18 | 7.73 |
| compound-beta-mini | 0.55 | 0.02 | 26.83 | 0.48 | 0.43 | 0.21 | 4.18 |
| google/flan-t5-small | 0.78 | 0.03 | 55.97 | 0.05 | 0.40 | 0.47 | 1.19 |
| google/flan-t5-base | 0.83 | 0.03 | 60.94 | 0.00 | 0.31 | 0.20 | 0.82 |
| google/flan-t5-large | 0.80 | 0.02 | 41.78 | 0.14 | 0.50 | 0.88 | 4.37 |

All metrics were normalized to the range $[0, 1]$ to ensure comparability. Moreover, to harmonize the graphical interpretation, all metrics were transformed so that higher values consistently reflected better performance. In particular, the generation time metric was inverted after normalization to indicate a preference for faster models: in the radar charts, higher scores on the *Time* dimension correspond to quicker and thus more desirable inference. Finally, the use of color coding in the legend groups together LLMs belonging to the same family, enabling a more immediate comparative analysis.

### 3.1. Quantitative Evaluation of ANFIS Classification

Although the primary goal of this work is not to assess the predictive capability of the classifier, we report the quantitative classification results to confirm that the model adequately captures the underlying data distribution, even on the test set. This ensures that the derived IF–THEN explanations are consistent with the information contained in the data. A data split was applied to guarantee a balanced representation of the classes and to obtain a more reliable performance evaluation. Specifically, the original dataset was partitioned into two subsets: 70% of the data was used for training, while the remaining 30% was reserved for testing. On the test set, the model achieved precision, recall, and F1-scores of 0.95 for both classes, along with an overall accuracy of 0.95, indicating that the classifier provides a faithful representation of the data.

### 3.2. Quantitative Evaluation of Explanations with the Zero-Shot Prompting Pattern

We conducted a detailed analysis of the generated explanations by examining each quality dimension individually. Table 1 reports the results obtained with the zero-shot pattern. With respect to lexical diversity, the Flan-T5 models produced the richest and most varied vocabulary. However, this apparent advantage is offset by their poor coherence, revealing a disconnect between lexical richness and logical structure. In contrast, instruction-tuned LLaMA variants offered a more balanced profile, striking a balance between acceptable diversity and greater structural consistency. In terms of readability, Flan-T5 models again stand out, generating text that is particularly simple and easy to follow. However, this comes at the cost of lower semantic grounding, as reflected in their weak coherence. Many LLaMA and Compound models instead yielded lower readability, likely due to their tendency toward more technical or formal language. Coherence, understood as the logical continuity of the explanation and its alignment with the prompt, was strongest in the Compound models and some LLaMA variants, which consistently produced semantically consistent and well-structured content. Flan-T5 models, by contrast, consistently underperformed in this respect. Coverage of information, assessed through both embedding-based and token-based measures, was highest for models such as Compound and LLaMA-4 variants, which better

**Table 2**
Quantitative evaluation of explanations generated with the persona pattern prompt across different LLMs.

| Model | TTR | Maas | FRES | CohS | ECS | TCS | Time |
|---|---|---|---|---|---|---|---|
| llama3-8b-8192 | 0.53 | 0.02 | 20.26 | 0.50 | 0.39 | 0.25 | 1.44 |
| llama-3.1-8b-instant | 0.57 | 0.01 | 15.83 | 0.49 | 0.38 | 0.24 | 1.42 |
| llama-3.3-70b-versatile | 0.62 | 0.01 | 12.72 | 0.54 | 0.39 | 0.27 | 2.04 |
| llama3-70b-8192 | 0.60 | 0.01 | 20.02 | 0.52 | 0.40 | 0.25 | 1.98 |
| allam-2-7b | 0.59 | 0.01 | 19.10 | 0.40 | 0.40 | 0.23 | 0.78 |
| deepseek-r1-distill-llama-70b | 0.50 | 0.01 | 36.03 | 0.41 | 0.35 | 0.17 | 3.49 |
| meta-llama/llama-4-maverick-17b-128e-instruct | 0.62 | 0.01 | 16.41 | 0.54 | 0.38 | 0.25 | 2.59 |
| meta-llama/llama-4-scout-17b-16e-instruct | 0.59 | 0.01 | 17.63 | 0.54 | 0.36 | 0.22 | 1.88 |
| qwen-qwq-32b | 0.55 | 0.01 | 30.17 | 0.38 | 0.34 | 0.14 | 3.88 |
| compound-beta | 0.57 | 0.01 | 14.36 | 0.53 | 0.39 | 0.24 | 5.76 |
| compound-beta-mini | 0.58 | 0.01 | 17.15 | 0.54 | 0.36 | 0.24 | 3.88 |
| google/flan-t5-small | 0.91 | < 0.01 | 55.52 | 0.00 | 0.30 | 0.20 | 1.02 |
| google/flan-t5-base | 0.78 | 0.03 | 52.65 | 0.08 | 0.29 | 0.18 | 2.46 |
| google/flan-t5-large | 0.82 | 0.03 | 48.52 | 0.00 | 0.29 | 0.11 | 2.35 |

integrated relevant concepts into their generated output. While Flan-T5 models achieved high token overlap, this often reflected surface-level repetition rather than substantive incorporation of information. Finally, the analysis of inference times highlights clear differences in computational efficiency. Larger models generally required longer generation times, confirming the correlation between parameter size and latency. Nonetheless, some optimized LLaMA models achieved relatively low inference times despite their scale, indicating that architectural refinements and implementation strategies play a significant role. Smaller models, as expected, remained the fastest, suggesting their suitability for real-time applications.

Taken together, the results show that no single LLM clearly dominates across all evaluation dimensions. Flan-T5 models excel in readability and lexical richness but struggle with coherence. Compound and LLaMA models achieve stronger coherence and information coverage, though sometimes at the cost of readability or efficiency. Overall, LLaMA-based models appear to offer the most balanced trade-off, whereas the optimal choice depends on which explanatory dimension is prioritized for the application context.

### 3.3. Quantitative Evaluation of Explanations with the Persona Prompting Pattern

Table 2 reports the quantitative results obtained with the persona pattern. The analysis reveals that Flan-T5 models exhibit very high lexical diversity and readability; however, this comes at the expense of extremely low coherence and limited coverage, which undermines the overall reliability of their explanations. In contrast, the LLaMA family, including both medium- and large-scale variants, achieves a more balanced profile: their lexical diversity is slightly lower, but they compensate with stronger coherence and more stable information coverage, even though their readability remains modest. The Compound models behave similarly, offering good coherence and coverage but with longer generation times. Qwen and Deepseek distinguish themselves for higher readability compared to most LLaMA models, although they lag in coherence and coverage.

Taken together, these results confirm that no single LLM excels across all dimensions: Flan-T5 stands out only for readability and diversity. At the same time, LLaMA- and Compound-based models appear better suited for generating coherent and semantically grounded explanations, even if their readability is less immediate. Overall, the persona pattern favors models with balanced behavior, suggesting that LLaMA variants offer the most stable trade-off.

**Table 3**
Quantitative evaluation of explanations generated with the persona pattern and fact-checking prompts across different LLMs.

| Model | TTR | Maas | FRES | CohS | ECS | TCS | Time |
|---|---|---|---|---|---|---|---|
| llama3-8b-8192 | 0.53 | 0.01 | 21.72 | 0.36 | 0.39 | 0.21 | 2.14 |
| llama-3.1-8b-instant | 0.52 | 0.01 | 26.22 | 0.31 | 0.39 | 0.24 | 1.50 |
| llama-3.3-70b-versatile | 0.60 | 0.01 | 11.06 | 0.51 | 0.37 | 0.27 | 2.09 |
| llama3-70b-8192 | 0.53 | 0.02 | 19.77 | 0.43 | 0.38 | 0.23 | 2.31 |
| allam-2-7b | 0.48 | 0.02 | 27.02 | 0.30 | 0.37 | 0.23 | 0.84 |
| deepseek-r1-distill-llama-70b | 0.47 | 0.01 | 33.03 | 0.33 | 0.33 | 0.18 | 4.25 |
| meta-llama/llama-4-maverick-17b-128e-instruct | 0.57 | 0.01 | 15.80 | 0.39 | 0.36 | 0.28 | 2.29 |
| meta-llama/llama-4-scout-17b-16e-instruct | 0.54 | 0.01 | 15.40 | 0.49 | 0.36 | 0.23 | 1.76 |
| qwen-qwq-32b | 0.52 | 0.01 | 37.19 | 0.37 | 0.33 | 0.16 | 5.67 |
| compound-beta | 0.50 | 0.02 | 13.99 | 0.48 | 0.38 | 0.26 | 6.56 |
| compound-beta-mini | 0.54 | 0.01 | 20.19 | 0.47 | 0.36 | 0.24 | 4.42 |
| google/flan-t5-small | 0.91 | < 0.01 | 44.32 | 0.28 | 0.38 | 0.28 | 1.20 |
| google/flan-t5-base | 0.77 | 0.04 | 50.94 | 0.07 | 0.29 | 0.17 | 2.41 |
| google/flan-t5-large | 0.79 | 0.03 | 46.44 | 0.19 | 0.31 | 0.16 | 3.51 |

## 3.4. Quantitative Evaluation of Explanations with the Persona and Fact-checking Prompting Pattern

Table 3 presents the results obtained by combining the persona pattern with fact-checking. The analysis shows that Flan-T5 models continue to achieve the highest lexical diversity and readability. However, their performance is undermined by very low coherence and limited coverage, which reduces the reliability of the generated explanations. LLaMA-based models, particularly the larger variants, offer a more balanced profile, maintaining moderate lexical diversity and readability while achieving stronger coherence and stable coverage of information. The Compound models also perform well in terms of coherence and coverage, but are penalized by slower inference times. Qwen and Deepseek display higher readability than most LLaMA variants, though their coherence and coverage remain weaker.

Overall, no single model clearly dominates across all dimensions; Flan-T5 stands out only in terms of diversity and readability. In contrast, LLaMA and Compound models provide more coherent and semantically grounded explanations, confirming their suitability when factual accuracy and logical consistency are prioritized.

## 3.5. Quantitative Comparison of Explanations Obtained with Different Prompt Patterns

Observing the radar charts in Figure 4, a clear difference emerges between the results obtained with the zero-shot pattern (4a) and those based on the persona pattern (4b and 4c). In the zero-shot setting, Flan-T5-large (light violet line) stands out for lexical diversity and coverage, suggesting that these models generate explanations with richer vocabulary and broader reference to the content of the prompt. Flan-T5-small and Flan-T5-base (dark violet lines) excel in readability, producing outputs that are simple and easy to follow, although this fluency does not necessarily translate into depth or precision of reasoning. By contrast, LLaMA3-8B-8192 (light blue line) achieves the best results in coherence and generation time, indicating a stronger ability to preserve logical flow across sentences while also being computationally efficient. The remaining LLMs are concentrated mainly around the coherence dimension, reflecting an effort to maintain internal consistency; however, they perform more modestly in terms of coverage, readability, and diversity, which limits the richness and accessibility of their explanations.

With the persona pattern, however, the distribution of LLMs in the radar chart changes substantially, supporting the hypothesis that different prompt patterns emphasize different explanatory qualities. Flan-T5-small (dark violet line) emerges with strong performance in time, diversity, readability, and

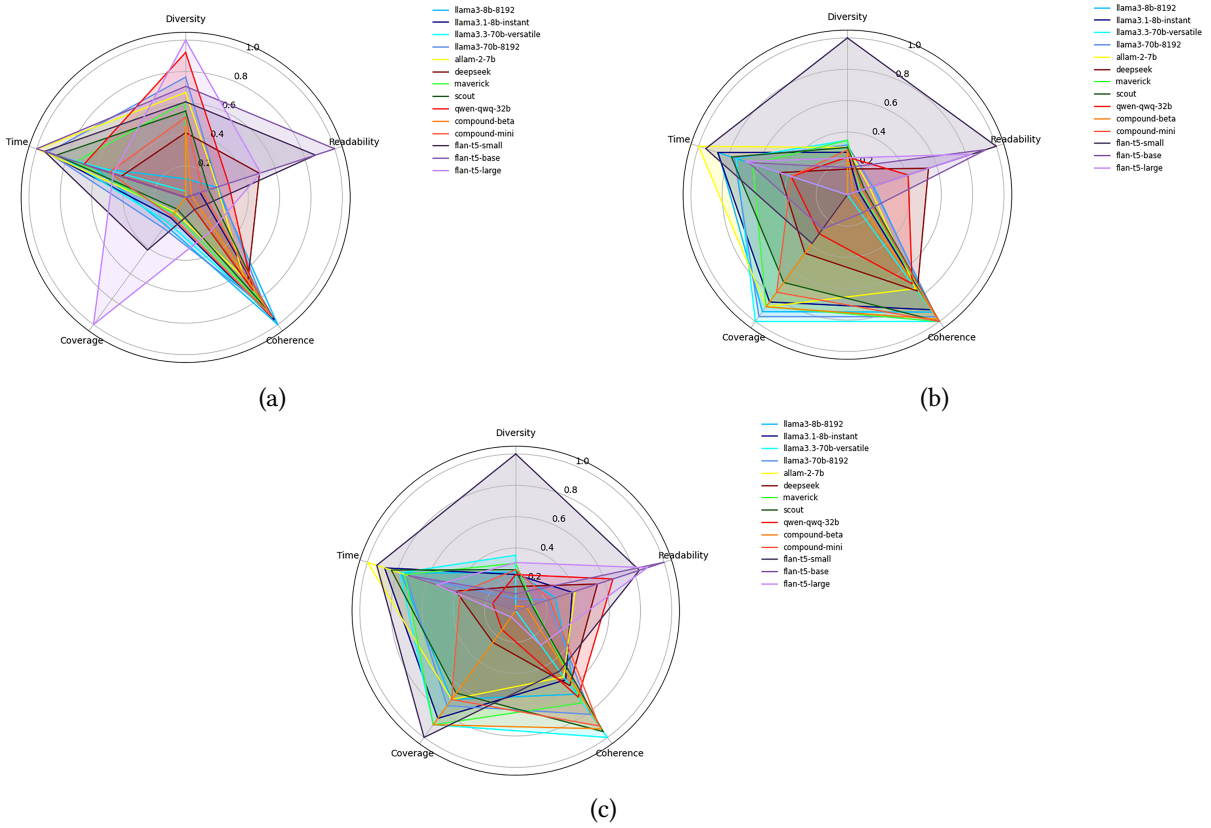



**Figure 4:** Radar charts summarizing the results for the 14 LLMs across the five dimensions, Diversity, Readability, Coherence, Coverage, and Time, using (a) the zero-shot pattern, (b) the persona pattern, and (c) the persona pattern with fact-checking.

partially in coverage, showing that it can generate accessible and lexically rich explanations with good alignment to the prompt, though still lacking in coherence. Most of the other LLMs cluster around coherence and coverage, prioritizing structured and content-grounded explanations, albeit at the expense of readability and lexical variety. LLaMA3-8B-8192 (light blue line) once again shows the highest values.

When comparing the two radars based on the persona and persona+fact-checking patterns (Figures 4b and 4c), the distributions appear largely similar. However, it is noteworthy that Flan-T5-small (dark violet line) now spans almost all dimensions, achieving high or near-maximum values in every aspect except coherence, suggesting that fact-checking reinforces its strengths in diversity, readability, and coverage without fully remedying its weakness in logical continuity.

## 3.6. Qualitative evaluations of explanations

While the quantitative metrics outlined above provide an objective and comparable assessment of model performance, they do not fully capture the nuances of linguistic quality or the practical usability of the generated explanations. To complement and validate the numerical results, a focused qualitative analysis was conducted to gain a deeper understanding of the observed behaviors. Given the extensive amount of data, it was not feasible to examine all prompts and responses exhaustively; instead, the analysis concentrated on the most representative cases identified through the quantitative evaluation.

Figures 5a and 5b illustrate an example of a zero-shot prompt for a non-epileptic and an epileptic case, respectively. These examples will serve as references for analyzing the outputs of different models. Please note that we do not report the prompts generated with the other patterns, since they follow the same rule presented here and rely on the templates already introduced.

Figure 6 shows three examples of explanations generated for both epileptic and non-epileptic cases in

```
The system has determined that the patient is in a state of NON-EPILEPTIC.
        This result was obtained using a fuzzy rule-based inference system.
        The rules that best explain the patient's condition are as follows:
        - IF DELTA is MEDIUM and THETA is MEDIUM and ALPHA is HIGH and BETA is MEDIUM and GAMMA is MEDIUM THEN NON-EPILEPTIC
        Should further details or clarification be required, the system can provide a more comprehensive explanation.
        Please, be discursive and do not use lists.
```

(a)

```
The system has determined that the patient is in a state of EPILEPTIC.
        This result was obtained using a fuzzy rule-based inference system.
        The rules that best explain the patient's condition are as follows:
        - IF DELTA is HIGH and THETA is HIGH and ALPHA is HIGH and BETA is MEDIUM and GAMMA is MEDIUM THEN EPILEPTIC
        Should further details or clarification be required, the system can provide a more comprehensive explanation.
        Please, be discursive and do not use lists.
...
```

(b)

**Figure 5:** Examples of prompts with the zero-shot pattern for (a) a sample predicted as epileptic and (b) a sample predicted as non-epileptic, together with the fuzzy rule derived from ANFIS.

a zero-shot setting using three different LLMs. Figures 6a and 6b illustrate the explanations generated for the non-epileptic and epileptic cases using LLaMA3-8b-8192 in a zero-shot learning setting. The explanations are organized into four sections: the first reports the decision produced by the fuzzy rule-based system, clarifying its meaning and context of application; the second provides a textual description of the rule that led to this decision; the third expands on this rule by analyzing the frequency band values, their respective intensities, and their relationship with the patient's states, thereby demonstrating the model's ability to exploit the knowledge embedded in its training; and the fourth concludes with a comprehensive summary of the reasoning process, emphasizing the key elements of the inferential outcome. This is consistent with the quantitative evaluation, which indicated a high degree of coherence.

Figures 6c and 6d show the explanations generated by Flan-T5-base. Despite the quantitative metrics indicating high readability, elevated lexical diversity, and null coherence, the explanations consist merely of repetitions of the fuzzy rules. The LLM neither contributes additional knowledge nor structures the explanation as required, resulting in poor and uninformative outputs. This excessive brevity and lack of narrative connection not only confirm the low coherence but also reveal that the observed lexical diversity stems from the use of non-repetitive technical terms rather than from genuine richness and variety of prose. Such overly concise and poor explanations highlight the model's limitations in producing meaningful interpretative content.

Finally, Figures 6e and 6f illustrate the explanations obtained with DeepSeek-R1-Distill-LLaMA-70B. In this case, the average values of the five quantitative dimensions shown in the radar chart are confirmed by the qualitative analysis. The explanations are very lengthy and include an initial informal part, where the model explicitly verbalizes the reasoning process it should follow to generate the explanation. This introduces redundancy and confusion, ultimately weakening the interpretability of the output. In the actual explanation, outside the 'think' tags, the fuzzy rule itself is not reported; however, references to the frequency bands and their connection to the predicted state are included, drawing on the knowledge embedded in the LLM, along with a mention of the fuzzy inference system. Without the preliminary reasoning section, the explanations would be considerably more effective. Notably, in the epileptic case, the model digresses by providing a general description of fuzzy logic itself.

From this initial analysis, it emerges that while quantitative metrics help measure the quality of explanations, they do not always align with our perception of what constitutes a "good" explanation, as exemplified by the case of Flan-T5-base.

Figure 7 presents two examples of LLM-generated explanations for non-epileptic and epileptic cases. We report the explanations obtained with LLaMA3-8b-8192 in Figures 7a and 7b, since the radar chart showed that, in addition to maintaining coherence in the zero-shot setting, this model also gains coverage when adopting the persona pattern. The structure of the explanations, however, varies between cases. In the non-epileptic explanation, the model first reports the rule, then describes the

Based on the analysis, the system has determined that the patient is in a state of non-epileptic condition. This conclusion was reached through a fuzzy rule-based inference system, which evaluated various factors to arrive at this outcome.

The key factors that contributed to this determination were the medium levels of delta, theta, and beta waves, as well as the high level of alpha waves and medium level of gamma waves. The combination of these factors led the system to infer that the patient's condition is non-epileptic.

In other words, the presence of medium levels of delta and theta waves, which are typically associated with drowsiness and relaxation, respectively, suggests that the patient is not experiencing an epileptic seizure. The high level of alpha waves, which are often seen in states of relaxation and closed eyes, also supports this conclusion. Meanwhile, the medium levels of beta and gamma waves, which are typically associated with attention and cognitive processing, respectively, indicate that the patient's brain activity is not excessively abnormal.

Overall, the system's analysis suggests that the patient's condition is not epileptic in nature, but rather a normal or abnormal condition that is not related to epilepsy. If further clarification or details are required, the system can provide a more comprehensive explanation of the analysis and the factors that contributed to this conclusion.

(a)

Based on the analysis, the system has determined that the patient is in a state of epilepsy. This conclusion was reached through the application of a fuzzy rule-based inference system, which considers various factors related to brain activity.

The system's rules indicate that the patient's condition is best explained by a combination of high levels of delta, theta, alpha, and beta brain waves, along with medium levels of gamma brain waves. This specific pattern of brain activity suggests that the patient is experiencing a state of epilepsy.

In particular, the presence of high levels of delta, theta, alpha, and beta brain waves indicates that the patient's brain is experiencing a heightened state of activity, which can be indicative of an abnormal electrical discharge. The medium levels of gamma brain waves, on the other hand, suggest that the patient's brain is not entirely synchronized, which can also contribute to the development of an epileptic episode.

Overall, the system's analysis suggests that the patient's brain activity is characterized by a complex interplay of different brain waves, which collectively contribute to the development of an epileptic state.

(b)

NON-EPILEPTIC:

if DELTA is MEDIUM and theTA is MEDIUM and ALPHA is HIGH and BETA is MEDIUM and GAMMA is MEDIUM then NON-EPILEPTIC.

(c)

EPILEPTIC:

The patient is in a state of EPILEPTIC.

(d)

<think>
Okay, so I need to figure out how to explain the system's determination that the patient is in a non-epileptic state. The system used a fuzzy rule-based inference system, which I remember is a type of fuzzy logic used for decision-making.

The main rule that applies here is: IF DELTA is MEDIUM and THETA is MEDIUM and ALPHA is HIGH and BETA is MEDIUM and GAMMA is MEDIUM THEN NON-EPILEPTIC.

Hmm, I should start by understanding what each of these bands (delta, theta, alpha, beta, gamma) represents. Delta waves are usually associated with deep sleep or unconsciousness, theta with drowsiness or early sleep stages, alpha with relaxation or closed eyes, beta with active thinking or concentration, and gamma with higher mental activity like problem-solving.

So the patient has medium delta, which might mean they're not in a deep sleep or unconscious. Theta is medium too, so maybe they're not drowsy or in the early stages of sleep. Alpha being high is interesting because that's typically a relaxed state, maybe the patient is calm. Beta and gamma are medium, so they might not be actively thinking or engaged in complex tasks.

Putting this together, the combination of these medium and high values doesn't fit the typical epileptic patterns. In epilepsy, especially during seizures, there might be more pronounced theta or delta activity, or maybe higher gamma activity due to hyperactivity. So the system is seeing a more balanced and calm state, leading to the conclusion of non-epileptic.

I should explain this in a way that's easy to understand, without using lists. Maybe start by introducing the system, then go into each band's role, how their levels indicate a non-epileptic state, and what that means for the patient.

Wait, I should make sure I'm accurate about the wave bands. Let me double-check: delta is low frequency, associated with sleep; theta with drowsiness; alpha with relaxation; beta with active, engaged states; gamma with higher cognitive functions. So medium levels in delta and theta suggest the patient isn't in a deep sleep or drowsy, which could be signs of a different condition.

High alpha suggests relaxation, which is common when a person is awake but calm. Medium beta and gamma might mean the patient isn't actively thinking or processing information intensely, which could be different from someone experiencing an epileptic seizure, where brain activity might be more erratic or heightened in certain bands.

I think I've got a good grasp now. I'll structure the explanation by first introducing the system, then explain each band's level and what it signifies, and conclude with how these together lead to the non-epileptic determination.
</think>

The system has determined that the patient is in a non-epileptic state based on an analysis of brain wave activity using a fuzzy rule-based inference system. This determination was made by evaluating the levels of different frequency bands in the patient's brain activity: delta, theta, alpha, beta, and gamma.

Delta waves, which are typically associated with deep sleep or unconsciousness, were found to be at a medium level. This suggests that the patient is not in a state of deep sleep or unconsciousness. Theta waves, linked to drowsiness or early stages of sleep, were also medium, indicating that the patient is not in a drowsy state or in the early stages of sleep.

The alpha wave level was high, which is usually associated with a relaxed yet alert state, such as when a person is calm and relaxed with their eyes closed. Beta waves, which are linked to active thinking, concentration, and engagement, were at a medium level, suggesting that the patient is not actively thinking or concentrating intensely. Gamma waves, associated with higher cognitive functions like problem-solving, were also medium, indicating that the patient is not engaged in complex mental tasks.

The combination of these medium and high levels across the different frequency bands does not align with the typical patterns observed in epileptic states. Epileptic activity often presents with more pronounced theta or delta wave activity, or heightened gamma activity due to hyperactivity in the brain. In contrast, the patient's brain activity appears more balanced and calm, leading to the conclusion that the patient is in a non-epileptic state.

This determination is based on the fuzzy rule-based system's evaluation of the patient's brain wave patterns, which suggests a state of relaxation and normal brain activity rather than the hyperactivity or irregular patterns typically associated with epilepsy.

(e)

<think>
Okay, so I need to figure out how to respond to this query about a patient being in an epileptic state based on some fuzzy rules. Let me start by understanding the given information. The system used a fuzzy rule-based inference system and identified that the patient is epileptic. The rule that fired was: IF DELTA is HIGH and THETA is HIGH and ALPHA is HIGH and BETA is MEDIUM and GAMMA is MEDIUM THEN EPILEPTIC.

Hmm, the user wants a discursive explanation without using lists. So I should probably break down each part of the rule and explain how each band contributes to the diagnosis. But first, I should explain what fuzzy logic is in simple terms because not everyone might be familiar with it.

I remember that in EEGs, different frequency bands are associated with different states. Delta waves are usually linked to unconsciousness or deep sleep, theta to drowsiness, alpha to relaxation, beta to active thinking, and gamma to higher cognitive functions. So, high delta and theta might indicate some abnormality since they're not typically high when someone is awake and relaxed.

The rule has all these bands set to specific levels. So, when delta, theta, and alpha are high, and beta and gamma are medium, it triggers the epileptic conclusion. I should explain each band's role and how their combination points to epilepsy. Maybe also mention that fuzzy logic allows for degrees of membership, so it's not just a yes or no but a matter of degree.

I should also make it clear that this is a simplified explanation and that actual diagnosis requires more comprehensive analysis. The user might be a medical professional looking for a concise yet thorough explanation, or perhaps someone less familiar who needs the basics laid out without jargon.

I need to structure the response logically: start with what epilepsy is, then explain the EEG bands, then go into fuzzy logic, and finally tie it all together with the specific rule. I should avoid technical terms where possible or explain them when necessary.

Wait, I should also consider that the user might be looking for reassurance that the system's diagnosis is reliable. So, I can mention that fuzzy logic is a robust method for handling the uncertainty and imprecision in biological data, which is common in medical diagnostics.

I should also touch on the implications of each band being in a certain state. For example, high delta and theta might suggest some sort of dysregulation, while medium beta and gamma could indicate that the brain's normal activity is altered but not completely suppressed.

Putting it all together, the response should flow from general to specific, explaining the concepts and then applying them to the rule provided. I need to make sure it's clear, concise, and covers all the necessary points without being too technical.

I think I've covered the main points. Now, I'll structure the response to address each part step by step, ensuring it's discursive and informative without using lists.
</think>

The system has determined that the patient is in a state of epilepsy based on a fuzzy rule-based inference system. This conclusion was reached by evaluating the patient's EEG (electroencephalogram) data, which measures electrical activity in the brain. The EEG data is categorized into different frequency bands—delta, theta, alpha, beta, and gamma—each of which corresponds to specific states of brain activity.

In this case, the system identified that the patient's EEG data exhibited high levels of delta and theta waves, high alpha waves, and medium levels of beta and gamma waves. These findings are significant because they align with a predefined fuzzy rule that associates such a combination of EEG patterns with epilepsy. The rule states: "IF DELTA is HIGH and THETA is HIGH and ALPHA is HIGH and BETA is MEDIUM and GAMMA is MEDIUM THEN EPILEPTIC."

Fuzzy logic is a mathematical approach that allows for the use of linguistic variables and fuzzy sets to model uncertainty and imprecision in complex systems. Unlike traditional binary logic, which relies on crisp boundaries, fuzzy logic can handle the gradual and overlapping nature of biological data, making it particularly suitable for medical diagnostics. In this context, the system evaluates the degree to which each EEG band matches the linguistic terms (e.g., HIGH, MEDIUM) and combines these degrees of membership using fuzzy operators to determine the overall likelihood of epilepsy.

The high delta and theta waves in the patient's EEG suggest abnormal brain activity, as these frequencies are typically associated with states of unconsciousness, deep sleep, or drowsiness. Elevated alpha waves, which are usually linked to relaxation or closed eyes, further contribute to the abnormal pattern. Meanwhile, medium levels of beta and gamma waves indicate that the brain's normal waking activity and higher cognitive functions are somewhat diminished but not entirely absent.

(f)

**Figure 6:** From top to bottom and left to right: explanations generated in zero-shot mode for a non-epileptic case and an epileptic case using LLaMA3-8B-8192 (a), (b), Flan-T5-base (c), (d), and DeepSeek-R1-Distill-LLaMA-70B (e), (f).

relationships between the frequency band values and the output, and even provides details on the functioning of the fuzzy inference system. In the epileptic explanation, the structure changes: the rule is not explicitly reported, and the focus shifts to a detailed account of the band values and their relation to the predicted condition, with only a brief reference to the fuzzy inference system.

We did not report the explanations generated by Flan-T5; however, despite the radar chart indicating very high values of diversity and readability, as in the previous case the explanations are extremely

As we review the output of the fuzzy rule-based diagnostic model, it's clear that the system has identified a pattern in the patient's EEG data that is consistent with a non-epileptic condition. Specifically, the activation of the rule "IF DELTA is MEDIUM and THETA is MEDIUM and ALPHA is HIGH and BETA is MEDIUM and GAMMA is MEDIUM THEN NON-EPILEPTIC" suggests that the patient is not currently experiencing an epileptic seizure.

From a clinical EEG perspective, this pattern is notable because it indicates a relatively normal and balanced distribution of brain activity across different frequency bands. The medium levels of delta and theta activity suggest that the patient's brain is not overly slow or sluggish, which is often a characteristic of epileptic activity. The high level of alpha activity, on the other hand, indicates that the patient's brain is in a state of relaxed alertness, which is consistent with a non-epileptic condition.

The medium levels of beta and gamma activity also contribute to this conclusion, as they suggest that the patient's brain is not overly active or hyperaroused, which can also be a hallmark of epilepsy. Overall, this pattern of activity is more typical of a normal, non-epileptic brain, and the fuzzy rule-based system has correctly identified this pattern as indicative of a non-epileptic condition.

One of the strengths of this fuzzy inference system is its ability to mimic expert reasoning while handling uncertainty. By combining multiple EEG frequency band features into a single rule, the system is able to capture the complex and nuanced relationships between these features that are characteristic of clinical EEG interpretation. This allows the system to make accurate predictions even in the presence of noise or variability in the data, which is a common challenge in EEG analysis.

In this case, the system's ability to handle uncertainty is particularly important, as the distinction between epileptic and non-epileptic activity can be subtle and context-dependent. By incorporating multiple features and using fuzzy logic to combine them, the system is able to make a more informed and accurate diagnosis, even in the face of uncertainty.

(a)

The fuzzy rule-based diagnostic model has yielded a diagnosis of EPILEPTIC condition, driven by the activation of a specific rule. This rule, in essence, combines the presence of high amplitudes in multiple EEG frequency bands, including delta, theta, alpha, beta, and gamma. From a clinical EEG perspective, this pattern is indicative of an epileptic profile.

The delta band, typically ranging from 0.5 to 4 Hz, is often associated with slow-wave activity, which can be seen in various neurological conditions, including epilepsy. The presence of high amplitude delta waves in this rule suggests a heightened state of cortical activity, potentially indicative of an underlying epileptogenic process. The theta band, spanning 4-8 Hz, is also implicated in this rule. Theta activity is commonly observed in states of relaxation, sleep, and even certain neurological disorders. In the context of epilepsy, theta waves can be seen as a precursor to seizure activity or as a manifestation of the abnormal neuronal activity that underlies the condition. The alpha band, typically ranging from 8-12 Hz, is often associated with closed eyes, relaxation, and decreased cortical activity. However, in the context of epilepsy, alpha activity can be seen as a sign of cortical hyperexcitability, potentially contributing to the development of seizure activity. The beta band, spanning 13-30 Hz, is often linked to cognitive processes, such as attention and motor activity. In this rule, the presence of medium-amplitude beta waves suggests a heightened state of cortical activity, potentially indicative of an underlying epileptogenic process.

Finally, the gamma band, typically ranging from 30-100 Hz, is often associated with sensory processing, attention, and cognitive functions. Medium-amplitude gamma waves in this rule may indicate an abnormal synchronization of neural activity, potentially contributing to the development of seizure activity. The fuzzy inference system, in this case, effectively mimics expert reasoning by combining these frequency band features in a way that reflects the complex and dynamic nature of EEG patterns in epilepsy. By using fuzzy logic, the system can handle the inherent uncertainty and variability in EEG data, allowing it to make a diagnosis that is both accurate and robust. In essence, the activation of this rule suggests that the patient is exhibiting an EEG pattern that is characteristic of an epileptic condition. The combination of high amplitudes in multiple frequency bands, particularly delta, theta, and alpha, indicates a heightened state of cortical activity, potentially indicative of an underlying epileptogenic process. The fuzzy inference system has effectively distilled this complex pattern into a diagnosis of EPILEPTIC condition, providing a valuable tool for clinicians to aid in the diagnosis and management of this complex and challenging condition.

(b)

The fuzzy rule-based diagnostic model has determined that the patient is currently in a non-epileptic condition, based on the activation of a specific rule that takes into account the EEG frequency band features. This rule suggests that when delta, theta, and beta bands are medium, and alpha band is high, and gamma band is medium, the patient is likely in a non-epileptic state. From a clinical EEG perspective, this pattern reflects a relatively normal or stable cerebral activity profile. The medium levels of delta, theta, and beta bands indicate a balanced level of slow and fast wave activity, which is typically seen in individuals who are awake and relaxed. The high alpha band power is particularly noteworthy, as alpha waves are often associated with a state of relaxed yet alert wakefulness, characterized by closed eyes and reduced cortical activity. The presence of medium gamma band activity also supports this interpretation, as gamma waves are involved in higher-level cognitive processing and sensory perception. The combination of these frequency band features in this specific rule effectively captures a non-epileptic profile, as it deviates from the typical EEG patterns observed during epileptic seizures. Epileptic seizures often exhibit abnormal EEG patterns characterized by high-amplitude spikes, sharp waves, or spike-and-wave discharges, which are not evident in this case.

The fuzzy inference system used here mimics expert reasoning by incorporating the complex and nuanced relationships between EEG frequency band features and neurological conditions. By using fuzzy logic, the system can handle uncertainty and imprecision inherent in EEG interpretation, allowing for a more accurate and robust diagnosis. The system essentially replicates the way a neurologist would weigh the importance of different EEG features to arrive at a diagnosis, but does so in a more systematic and reproducible manner. In essence, the fuzzy rule-based model is emulating the clinical decision-making process of a neurologist, where the presence of a high alpha band power, combined with medium levels of delta, theta, beta, and gamma bands, leads to a conclusion that the patient is in a non-epileptic state. This approach enables the system to provide a clear and interpretable output that aligns with clinical expertise, while also acknowledging the inherent uncertainty and variability present in EEG data.

(c)

The fuzzy rule-based diagnostic model has identified the patient as being in an epileptic condition, driven by the concurrent activation of high delta, theta, and alpha power, alongside medium beta and gamma power. From a clinical EEG perspective, this pattern is highly suggestive of an epileptogenic process. In epilepsy, the EEG often exhibits abnormal frequency band activity, which can be characterized by an increase in lower frequency bands, such as delta and theta, and an alteration in the normal alpha, beta, and gamma band activity. The presence of high delta power typically indicates abnormal brain activity, often seen in conditions with altered consciousness or severe cortical dysfunction. High theta power frequently accompanies states of drowsiness, early sleep stages, or certain types of epilepsy, particularly those with a focal or generalized seizure onset. The elevated alpha power may seem counterintuitive; however, in the context of epilepsy, it can represent an abnormal alpha band activity, sometimes seen as an interictal or ictal pattern. The combination of high delta, theta, and alpha power, along with medium beta and gamma power, points towards a complex abnormal EEG profile. In epilepsy, beta and gamma bands often show alterations; however, their medium levels in this context might reflect a transitional state or the presence of focal abnormalities.

The fuzzy rule implemented here effectively mimics expert clinical reasoning by integrating multiple EEG features to arrive at a diagnosis. Fuzzy logic systems are particularly adept at handling the inherent uncertainty and imprecision present in clinical data, such as EEG readings. By defining the input variables (delta, theta, alpha, beta, gamma) in fuzzy terms (e.g., high, medium) and establishing a rule base that mirrors clinical expertise, the system can effectively synthesize complex information to produce a diagnostic output. In essence, the activation of this fuzzy rule reflects a synthesis of key EEG features that are commonly observed in epileptic conditions. The model's output underscores the value of fuzzy logic in clinical decision-making, particularly in handling the nuances and variability inherent in EEG interpretation. By quantifying and integrating these EEG features within a fuzzy framework, the system provides a powerful tool for supporting the diagnosis of epilepsy and similar conditions, ultimately enhancing clinical decision-making.

(d)

**Figure 7:** From top to bottom and left to right: explanations generated in persona pattern mode for a non-epileptic case and an epileptic case using LLaMA3-8B-8192 (a), (b), and Compound-Beta-Mini (c), (d).

short and often consist only of the fuzzy rule—or in some cases not even that (e.g., “The fuzzy rule is based on a combination of EEG frequency band features”). By contrast, we included the explanations obtained with Compound-Beta-Mini, since the radar chart shows a significant increase in both coverage and coherence compared to the zero-shot setting. As illustrated in Figures 7c and 7d, the explanations are structured into two sections. The first defines the output, describes the use of the fuzzy system, reports the rule, and explains the correlations between the frequency bands and the decision. The second provides a concise explanation of what a fuzzy inference system is and why it can be helpful in the medical domain. In both examples, we also observe a notable shift in tone, which becomes more formal and aligned with medical discourse.

Finally, Figure 8 reports two examples of explanations for the epileptic case generated using the


(a)


(b)

**Figure 8:** From top to bottom and left to right: explanations generated in persona pattern + fact-checking mode for an epileptic case using LLaMA3-8B-8192 (a), and Compound-Beta-Mini (b).

persona pattern with fact-checking, obtained with LLaMA3-8B-8192 and Compound-Beta-Mini. In both cases, we observe that the explanations conclude with bulleted lists that provide factual references to support the preceding statements. This strategy is intended to enhance expert trust in the analyses produced by automatic systems and, consequently, in the generated explanations.

For the sake of brevity, the explanations generated by Flan-T5 are not reported; however, in this case, the model reproduces the prompt without providing a list of references. This confirms that, despite the quantitative metrics suggesting optimal performance, the practical outcome is unsatisfactory. Similar issues were observed with DeepSeek, whose explanations again contain a section labeled ‘think’, and which is unable to generate a factual list. The same limitation applies to Qwen. LLaMA, on the other hand, does provide factual statements but fails to include references. Therefore, the integration of fact-checking with these LLMs does not effectively increase user trust in the generated explanations.

## 4. Conclusion

This study presented a workflow for generating structured textual explanations from IF–THEN rules produced by a neuro-fuzzy inference system, using Large Language Models as the explanatory layer. Applied to a case study on EEG-based seizure classification, the workflow enabled a systematic evaluation of 14 LLMs from different families and sizes under three prompting strategies: zero-shot, persona, and persona combined with fact-checking.

The results demonstrate that explanation quality cannot be fully captured by quantitative metrics alone. While Flan-T5 achieved high scores in diversity and readability, its explanations were superficial and uninformative, in contrast to LLaMA and Compound models, which offered more coherent and content-grounded outputs at the cost of readability or efficiency. Among the prompting strategies, the

persona pattern emerged as the most effective, improving both quantitative performance and qualitative richness. Fact-checking added a degree of transparency but revealed uneven support across different models.

Building on these findings, there is a clear need for novel quantitative metrics that go beyond surface-level linguistic indicators to more effectively capture coherence, informativeness, and contextual alignment. Equally important is the involvement of domain experts, both in assessing the explanations and in guiding the explanation process according to their needs, which is essential for the development of reliable, trustworthy, and human-centered explainable AI systems. Furthermore, when dealing with text generated by LLMs, an in-depth investigation of hallucinations is required to avoid the inclusion of false or misleading content—an aspect of particular relevance in the medical domain.

## Acknowledgments

Gi.C. and G.Z. acknowledge the support of the project FAIR – Future AI Research (PE00000013), Spoke 6 – Symbiotic AI (CUP H97G22000210007) under the NRRP MUR program funded by NextGenerationEU. Ga.C, Gi.C. and G.V, acknowledge the support of the PNRR TT project ARIAS / M.A.M.M.A. (CUP B533D22000980006), which partially funded this research within the project FAIR - Future AI Research (PE00000013), Spoke 6—Symbiotic AI (CUP H97G22000210007) under the NRRP MUR program funded by the NextGeneration EU. A Ph.D. fellowship funds A.G.V.'s research within the Italian "D.M. n. 630, April 24, 2024" – under the NRRP, Mission 4, Component 2, Investment 3.3 – Ph.D. project "Explainability of Artificial Intelligence systems for applications in the medical field", co-supported by Bristol-Myers Squibb s.r.l. (CUP B91I24000160007).

## Declaration on Generative AI

Generative AI tools, specifically OpenAI's ChatGPT and Grammarly, were used exclusively for grammar correction and language refinement. The authors conceived, wrote, and validated all content.

## References

[1] A. G. Valerio, K. Trufanova, S. de Benedictis, G. Vessio, G. Castellano, From segmentation to explanation: Generating textual reports from MRI with LLMs, Computer Methods and Programs in Biomedicine (2025) 108922.

[2] M. Cremaschi, D. Ditolve, C. Curcio, A. Panzeri, A. Spoto, A. Maurino, Decoding the mind: A RAG-LLM on ICD-11 for decision support in psychology, Expert Systems with Applications 279 (2025) 127191.

[3] D. Schicchi, C. Limongelli, V. Monteleone, D. Taibi, A closer look at ChatGPT's role in concept map generation for education, Interactive Learning Environments (2025) 1–21.

[4] P. Ardimento, M. Capuzzimati, G. Casalino, D. Schicchi, D. Taibi, A novel LLM-based classifier for predicting bug-fixing time in Bug Tracking Systems, Journal of Systems and Software (2025) 112569.

[5] L. Laraspata, F. Cardilli, G. Castellano, G. Vessio, Enhancing human capital management through GPT-driven questionnaire generation, in: Proceedings of the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI 2024) co-located with 23th International Conference of the Italian Association for Artificial Intelligence (AIxIA 2024), CEUR-WS. org, 2024.

[6] F. Gaber, M. Shaik, F. Allega, A. J. Bilecz, F. Busch, K. Goon, V. Franke, A. Akalin, Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis, npj Digital Medicine 8 (2025) 263.

[7] S. Shool, S. Adimi, R. S. Amleshi, E. Bitaraf, R. Golpira, M. Tara, A systematic review of large language model (LLM) evaluations in clinical medicine, BMC Medical Informatics and Decision Making (2025).

[8] T. Mirzaei, L. Amini, P. Esmaeilzadeh, Clinician voices on ethics of LLM integration in healthcare: a thematic analysis of ethical concerns and implications, BMC Medical Informatics and Decision Making (2024).
[9] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence, Information fusion 99 (2023) 101805.
[10] L. Nannini, J. Alonso-Moral, A. Catala, M. Lama, S. Barro, Operationalizing Explainable AI in the EU Regulatory Ecosystem, IEEE Intelligent Systems (2024).
[11] A. Bilal, D. Ebert, B. Lin, LLMs for Explainable AI: A Comprehensive Survey, arXiv.org (2025).
[12] F. Mumuni, A. Mumuni, Explainable artificial intelligence (XAI): from inherent explainability to large language models, arXiv.org (2025).
[13] E. Paraschou, I. Arapakis, S. Yfantidou, S. Macaluso, A. Vakali, Mind the XAI Gap: A Human-Centered LLM Framework for Democratizing Explainable AI, arXiv.org (2025).
[14] K. Nimala, C. Shieh, R. Nareshkumar, V. S. Murugan, Scalable and transparent mental health support via XAI-LLM, International Journal of Information Technology (2025).
[15] A. J. Maria, C. Castiello, M. Luis, C. Mencar, et al., Explainable fuzzy systems: Paving the way from interpretable fuzzy systems to explainable AI systems, Studies in Computational Intelligence 970 (2021) 1–253.
[16] S. Singh, Neuro-Fuzzy Architectures for Interpretable AI: A Comprehensive Survey and Research Outlook, Journal of Machine Learning Research 1 (2025) 11.
[17] G. Casalino, G. Castellano, A. G. Valerio, G. Vessio, G. Zaza, Enhancing the Explainability of Neuro-Fuzzy Systems with Large Language Models: A Case Study on EEG-Based Epileptic Seizure Classification, in: 2025 International Joint Conference on Neural Networks (IJCNN), IEEE, 2025.
[18] X. Kong, S. Liu, L. Zhu, Toward Human-centered XAI in Practice: A survey, Machine Intelligence Research 21 (2024) 740–770.
[19] M. L. Bangerter, G. Fenza, D. Furno, M. Gallo, V. Loia, C. Stanzione, I. You, A Hybrid Framework Integrating LLM and ANFIS for Explainable Fact-Checking, IEEE Transactions on Fuzzy Systems (2024).
[20] S. Singh, Are Large Language Models Good At Fuzzy Reasoning?, Proceedings of the 2024 7th International Conference on Computational Intelligence and Intelligent Systems (2024).
[21] A. H. Alamoodi, O. Zughoul, D. David, S. Garfan, D. Pamučar, A. S. Albahri, A. S. Albahri, S. Yussof, I. M. Sharaf, A Novel Evaluation Framework for Medical LLMs: Combining Fuzzy Logic and MCDM for Medical Relation and Clinical Concept Extraction., Journal of medical systems (2024).
[22] P. M. Perez-Ferreiro, A. Catala, A. Bugarin-Diz, J. M. Alonso-Moral, Generating trustworthy explanations with a language model enriched by fuzzy rule-based systems, in: 2025 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2025.
[23] D. Karaboga, E. Kaya, Adaptive network based fuzzy inference system (ANFIS) training approaches: a comprehensive survey, Artificial Intelligence Review 52 (2019) 2263–2293.
[24] B. Mesraoua, F. Brigo, S. Lattanzi, B. Abou-Khalil, H. Al Hail, A. A. Asadi-Pooya, Drug-resistant epilepsy: definition, pathophysiology, and management, Journal of the neurological sciences 452 (2023) 120766.
[25] K. Rasheed, A. Qayyum, J. Qadir, S. Sivathamboo, P. Kwan, L. Kuhlmann, T. O'Brien, A. Razi, Machine learning for predicting epileptic seizures using EEG signals: A review, IEEE reviews in biomedical engineering 14 (2020) 139–155.
[26] R. Andrzejak, K. Lehnertz, C. Rieke, F. Mormann, P. David, C. Elger, Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state [dataset], 2001.
[27] J.-S. R. Jang, C.-T. Sun, E. Mizutani, Neuro-Fuzzy and Soft Computing: a computational approach to learning and machine intelligence, (No Title) (1997).
[28] N. K. Kasabov, Foundations of neural networks, fuzzy systems, and knowledge engineering, Marcel Alencar, 1996.
[29] S. Schulhoff, M. Ilie, N. Balepur, K. Kahadze, A. Liu, C. Si, Y. Li, A. Gupta, H. Han, S. Schulhoff,

et al., The Prompt Report: A Systematic Survey of Prompting Techniques, arXiv preprint arXiv:2406.06608 (2024).
[30] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, D. C. Schmidt, A prompt pattern catalog to enhance prompt engineering with chatgpt, arXiv preprint arXiv:2302.11382 (2023).
[31] G. Aaron, et al., The llama 3 herd of models, 2024. arXiv:2407.21783.
[32] B. M Saiful, at al., ALLam: Large language models for arabic and english, in: The Thirteenth International Conference on Learning Representations, 2025.
[33] L. Aixin, et al., Deepseek-v3 technical report, 2025. arXiv:2412.19437.
[34] Meta, Llama 4 Maverick 17B-128E-Instruct-FP8, https://www.llama.com/docs/llama-4-maverick, 2025. Version: 17B active parameters, 400B total parameters, 128 experts.
[35] Meta, Llama 4: A New Foundation Model, https://www.llama.com/docs/llama-4-scout, 2025. Version: Scout (17B).
[36] Y. An, et al., Qwen3 technical report, 2025. arXiv:2505.09388.
[37] Groq, Compound Beta, a Compound AI System, https://console.groq.com/docs/compound/systems/compound-beta, 2025. Version: 2025-07-23 (Stable); Powered by Llama 4 Scout and Llama 3.3 70B.
[38] Groq, Compound Beta Mini, a Compound AI System, https://console.groq.com/docs/compound/systems/compound-beta-mini, 2025. Version: 2025-07-23 (Stable); Powered by Llama 4 Scout and Llama 3.3 70B.
[39] W. Hyung, et al., Scaling instruction-finetuned language models, 2022. arXiv:2210.11416.
[40] M. Honnibal, I. Montani, S. V. Landeghem, A. Boyd, spaCy: Industrial-strength Natural Language Processing in Python, 2020.