



Fusion of fuzzy rule-based factual and counterfactual explanations with large language models

Pablo Miguel Perez-Ferreiro ^{a,b,*}, Jose M. Alonso-Moral ^{a,b},
Alejandro Catala ^{a,b}, Alberto Bugarín-Diz ^{a,b}

^a Centro Singular de Investigación en Tecnoloxías Intelixentes (CITIUS), Universidade de Santiago de Compostela, 15782, Santiago de Compostela, Spain

^b Departamento de Electrónica e Computación, Universidade de Santiago de Compostela, 15782, Santiago de Compostela, Spain

ARTICLE INFO

Keywords:

Explainable AI
Counterfactuals
Large language models
Fuzzy rule-based systems

ABSTRACT

In this paper, we explore various strategies for generating trustworthy, natural narrative explanations through the combination of fuzzy rule-based systems with large language models. Specifically, we aimed to leverage contrastive, counterfactual explanations to supplement the factual pieces of information provided by the fuzzy systems to a template-based explainer; and to leverage Chain-of-Thought prompting schemes to improve the generation from language models. We have investigated the validity of the following two research hypotheses: (1) the addition of counterfactuals, which makes for a more complete explanation overall, does not significantly penalize trustworthiness; and (2) Chain-of-Thought prompting induces better discourse structuring and content planning on the language model side, so that it can effectively rewrite the template-based explanations into a higher-quality narrative, yet remaining a truthful explanation. We validated these research hypotheses in a use case in the medical field, comparing several metrics related to predictive performance, trust, brevity and clarity in explanation for several experimental setups. After experimentation, we found that the addition of counterfactual information is useful because it neither increases hallucinatory tendency nor decreases accuracy, hence not harming trustworthiness. However, Chain-of-Thought prompting does not induce the desired effects. On the contrary, it increases verbosity, and in consequence, it harms readability in the final narratives, with no apparent benefits.

1. Introduction

According to the latest Artificial Intelligence (AI) Index Report [1] by the Stanford Institute for Human-Centered AI: “the emergence of AI applications as everyday companions or assistants for jobs has sparked discussions about their implications, regarding practical but also ethical and legal terms”. On the one hand, the impact of AI soars in research or economic contexts with record investment and performance. On the other hand, public perception of trust in AI is not so high: while a positive attitude towards AI has grown in the last year, the share of surveyed individuals considering AI beneficial is barely over half.

The worldwide adoption of Deep Learning [2] as a predominant paradigm in AI, with “black box” opaque generative AI models in general and Large Language Models (LLMs) in particular standing out at the core of most intelligent applications, calls for further

* Corresponding author.

E-mail addresses: pablo.perez.ferreiro@usc.es (P.M. Perez-Ferreiro), josemaria.alonso.moral@usc.es (J.M. Alonso-Moral), alejandra.catala@usc.es (A. Catala), alberto.bugarin.diz@usc.es (A. Bugarín-Diz).

<https://doi.org/10.1016/j.ijar.2026.109828>

Received 9 January 2026; Received in revised form 23 April 2026; Accepted 17 September 2026

Available online 22 September 2026

0888-613X/© 2026 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

research efforts to endow such applications with transparency and explainability in compliance with regulation [3]. It is worth noting that the internal workings of LLMs remain unclear and difficult to explain despite advancements in Explainable AI (XAI) [4], a term introduced by DARPA [5] to define AI systems that are able to proactively explain themselves. Trustworthiness and reliability are only taken for granted with “white box” transparent-by-design approaches. However, the lack of explainability of opaque models translates into a lack of trustworthiness and reliability, as it is either difficult or risky to trust a system one can not understand.¹

HuggingFace currently hosts an LLM Safety Leaderboard,² which ranks LLMs according to the *DecodingTrust* safety index [6]. Few candidates score highly in the ranking, and most of the best-performing models are not open source, and thus fully under the control of their private owners. Meta’s LLaMA family is one of the few open models that stand out for both performance and safety. However, while the leaderboard includes several metrics related to fairness and reliability under adversarial conditions, it is completely devoid of any test related to hallucinatory tendencies, one of the most concerning traits of LLMs in the context of trustworthy AI. It also does not attempt to measure the adequacy of LLMs for XAI, which makes even the highest-ranking models of dubious validity in critical domains. This indicates that, while modern LLMs may be considered relatively safe from a technical and ethical standpoint, they remain largely unproven in terms of trustworthiness.

Fuzzy logic [7,8] is a classical AI paradigm with strong theoretical foundations and decades of successful applications [9,10]. Fuzzy rule-based systems (FRBSs) are transparent-by-design models [11] that represent either expert or data-grounded knowledge in the form of IF-THEN fuzzy rules with linguistic propositions that relate concepts in natural language, which is the basic form in which humans communicate and comprehend. Accordingly, a fuzzy rule can be understood as a local explanation for a modeled phenomenon.

This work builds upon our previous efforts in developing trustworthy explainers, in which we first proposed an architecture for an explanatory interface based on FRBSs, with enriched utterances through fusion of diverse explanatory inputs [12]. This architecture was then extended and validated to include LLM-based verbalization, taking advantage of LLMs’ high fluency and naturalness without a significant loss of trustworthiness [13]. It is worth noting that by using the fuzzy explainer as a data-grounded explanation provider, we successfully reduced the appearance of hallucinatory content.

In this work, we have expanded and enhanced both the FRBS explainer and the LLM-based verbalizer in several ways.

On the one hand, we have integrated counterfactual (or contrary to factual) explanations [14] alongside the already-tested, more common factual modality [15]. Counterfactual explanations (in essence, explanations of why something did not take place; thus, we can understand what should have been different to achieve another outcome) are believed to be a key part of human understanding, as a big portion of explanatory contexts have an implicitly contrastive nature [16]: when demanding an explanation, an individual is not only concerned as to why something happened, but also how the phenomena could have been prevented or altered.

On the other hand, we have enriched the text generator with Chain-of-Thought (CoT) prompting [17], an approach in which the models are instructed to perform a full reasoning chain about their task, and how they can fulfill it appropriately, before starting to “write” anything. This self-reinforcement, akin to a human breaking a complex task down into steps they can more easily understand, has been shown to be very helpful in a number of contexts. As discourse planning is a key step in ensuring a written piece is of high quality, performing a CoT pass may help the model structure its utterances better.

Hence, the novelty in this work is twofold: (i) integrating counterfactuals and CoT in the generation of narrative explanations; and (ii) verifying that the automated explanations are with a good reliability-trustworthiness trade-off.

The rest of the manuscript is structured as follows. Section 2 provides readers with an introduction to the state of the art and preliminary concepts related to the work. Section 3 details the proposal, describing both the general architecture of the explainer and an in-depth overview of its main components, along with an overview of the validation metrics and experimental setups designed to evaluate the system. Section 4 presents and discusses the results that were obtained, both quantitatively and qualitatively. Finally, Section 5 concludes with general remarks, and sketches out future research lines.

2. Related work

We begin by revisiting the state-of-the-art approaches for generating factual and counterfactual textual explanations (Section 2.1). Then, we introduce the most outstanding approaches for endowing LLMs with textual explainability (Section 2.2). We finish with a brief review of strategies for evaluating the quality of textual explanations (Section 2.3).

2.1. Factual and counterfactual explanations

An explanation may be defined as a presentation of the reasoning or behavior of an explicandum (i.e., whatever is being explained), from an explainer to an explainee [18]. This presentation may be partial, and in most contexts is meant to take place in human-understandable terms. Following this notion, it is reasonable that the most self-evident mode of explanation is one in which the rationale behind a certain outcome is directly exposed; e.g., when using a decision tree for classification, one may list the different data features that have led to the tree placing the data instance in a certain class [15]. We call this a “factual explanation”, as it concerns the facts pertaining to the data instance being classified (i.e., the explanation is characterized in terms of the actual features of the data instance).

¹ <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

² <https://huggingface.co/spaces/AI-Secure/llm-trustworthy-leaderboard> [Accessed by 11 December 2025]

This type of approach has spawned extensive literature. Pecori et al. [19] proposed a methodology for adapting the Hoeffding Decision Tree, a popular model in online learning, to the fuzzy paradigm, successfully improving its predictive power while simultaneously creating highly explainable trees. Idris and Ismail [20] proposed the Fuzzy-ID3 algorithm as an extension of the ID3 method for decision tree creation. Its use was demonstrated in breast cancer prediction, yielding models with a good balance between performance and explainability. The same use case was shared by [21], which explored the usage of crisp decision trees as sources of interpretable fuzzy rules. Ducange et al. [22] have also used fuzzy decision trees for building highly explainable classifiers in the healthcare domain. Namely, they showed that fuzzy decision trees can achieve performance comparable to state-of-the-art neural architectures (with additional explanatory capabilities) for brain tumor classification.

Additionally, social and cognitive sciences have long found that human understanding based on explanation is strongly contrastive in nature [23]. Supporters of this idea propose that, when asking “Why” questions, humans are not exclusively interested in the reasons behind an event they know has taken place, but on the reasons behind alternative possibilities not taking place.

Following this idea, it is possible to imagine alternate scenarios in which different circumstances would lead to different outcomes [14]. In cognitive science, it is usual to refer to the mental representations of such scenarios as “counterfactuals” (a shorthand for “contrary-to-fact”), as they describe circumstances that did not occur and thus explicitly contradict our factual knowledge about the world. Counterfactuals usually take the form of a conditional statement, in which an alternative to the factual series of events is described, and a consequence of the new situation is presented [24]. One could summarize counterfactuals as “If A, then C” statements, in which A represents a set of antecedents and C represents a consequent for them.

Counterfactuals have great interest in the field of XAI [25], where counterfactual explanations have improved the explanation quality of systems by virtue of their contrastive nature, which better imitates human thinking and incites a series of positive cognitive reactions in human users [26].

For instance, Tolomei et al. [27] introduced a method for extracting interpretable and actionable counterfactuals in tree-based ensemble models by finding the minimal tweaking of inputs features that turned true negatives into positives; validating their approach in online advertising quality evaluation. The idea of inducing perturbations on data instances to find sufficiently similar examples that dissent on the inferred consequent, and thus can be offered as counterfactuals, has also been explored by Guidotti et al. [15], who proposed a method for surrogate explanations of opaque models based in training local classifiers for specific instances, trained on a synthetic neighborhood, which were then used to extract both factual and counterfactual explanations. Stepin et al. [28] similarly proposed a method for counterfactual extraction in tree-based models, both crisp and fuzzy models were considered, automatically selecting the most relevant counterfactual from a wider set of potential explanations.

2.2. Textual explanations with LLMs

Following the breakthroughs enabled by the Transformer architecture [29], the field of Natural Language Processing (NLP) saw a rise in the performance levels achievable. While initial approaches, such as BERT [30] or GPT-1 [31], already showed great promise, surpassing state-of-the-art methods in many fields, they have been iterated on and improved over just a few years, achieving rapid progress.

Currently, state-of-the-art LLMs show outstanding performance for most NLP tasks [32,33]. Their fluency and similarity to human speech is compared favorably to other alternatives, and the generality of their training makes them capable of covering topics with impressive broadness. Perhaps most intriguingly, the progressive increasing of LLM sizing has prompted them to showcase what are called emergent abilities [34,35]: properties that were not accounted for during the design phase, and which are not present on smaller models and thus are not “predictable” by extrapolation.

Following that, there has been interest in developing strategies to improve the performance of existing LLMs on specific tasks, without modifying the models themselves. The *prompting* paradigm, introduced in [36], involves furnishing the LLM with a task instruction, written in plain natural language, to direct it towards the task’s completion (that instruction being the *prompt*). In addition, *few-shot prompting* involves several examples of a certain task’s adequate completion to guide the LLM in the processing of another instance of the problem.

A derivative of that approach is CoT prompting [17]. In this case, the model is given an input, its corresponding output, and the series of reasoning steps that would allow a human agent to traverse from the input to the output. This can be done with multiple examples, similar to few-shot approaches. The main advantage of the CoT approach over few-shot approaches is scalability, as few-shot learning often performs poorly on reasoning-heavy contexts. Nevertheless, CoT prompting works well across a range of complex tasks, even without providing elaborate examples. In [37], the authors showcased that induction of minimal CoT via prompting can improve LLM performance even on the extreme zero-shot (i.e., with no example given) learning context. A framework for automatically extracting suitable CoT rationales is presented in [38].

The achieved level of broad competence, along with their excellent capabilities for text generation, has drawn significant attention to LLMs as tools for explanation generation: as already discussed in the previous Section 2.1, explanations should be generated and conveyed in human-understandable terms, and natural language is the best medium for this. Rajani et al. [39], for instance, introduced a methodology to fine-tune BERT-based models to perform commonsense reasoning to state-of-the-art levels, while simultaneously producing sound explanations. In addition, Ma et al. [40] presented a model-agnostic framework for integrating LLMs with traditional collaborative filtering methods to produce explainable recommendation systems. Moreover, Lin et al. [41] leveraged CoT prompting in LLMs to transform tabular data into textual narratives that analyze and explain the original contents through a multi-stage pipeline, in which CoT traces are automatically sampled from LLMs, refined by a human, and fed back to the working LLM for generation of the final narrative.

In this vein, synergies between LLMs and fuzzy logic for explanatory tasks have been explored in the literature. Itoh and Kobayashi [42] leveraged fuzzy reasoning by fine-tuning a language model with natural language descriptions of the initial and final state of a physical object during a collision event. They automatically captured the ambiguous quantitative expressions in the narratives as fuzzy membership functions. Bangerter et al. [43] addressed a fact-checking task by combining LLMs and Adaptive Neuro-fuzzy Inference Systems: upon receiving a claim, the LLM decomposes it into subclaims, retrieves relevant evidence from vector storages or the Web, and utters a verdict and explanation; while simultaneously, the fuzzy system analyzes the relation between subclaims and retrieved evidence to produce a reliability score for the LLMs verdict.

Lastly, LLMs have also been employed in the context of counterfactual explanations. Gat et al. [44] proposed a dual methodology for counterfactual generation based on LLMs: either the LLM itself generates the counterfactual on inference time, or it is used to guide the training process of a problem-specific embedding space which can then be used to approximate counterfactuals. This methodology is applied as a method for explaining other opaque models. Fredes et al. [45], for their part, showcase a pipeline for counterfactual generation and verbalization based on LLMs, in which a set of diverse counterfactuals is generated using the DiCE framework [46] to then be analyzed and synthesized into a coherent narrative by a sequence of LLM calls.

2.3. Explanation evaluation

Explanations are fundamental to human cognition and understanding [47]. Not only do explanations provoke significant psychological effects, but they are also considered a guiding force in human reasoning. However, there is little consensus on what constitutes a good explanation in fundamental terms. While properties such as simplicity and broadness have been cited as key aspects leading to both adults and children perceiving explanations as good and satisfactory, it is likewise hard to define and quantify the impact of those two properties systematically [48].

In this vein, when evaluating explanations in social sciences, it is typical to do so via human-based studies (e.g., [49]). Nevertheless, in machine learning contexts, only some authors attempt to leverage human studies to distill relevant factors in creating high-quality explanations [50,51]. On the contrary, most publications go for automatic evaluation of explanatory quality (i.e., do not require a human study).

Readability [52], i.e., a multi-factorial property of a textual material that defines the success a group of readers will have with it, in terms of understanding, speed of completion, and enjoyment, can be used as a proxy for assessing the goodness of textual explanations. Readability has been extensively studied during the last century, in an attempt to increase readership and educational quality [53]. Many readability indexes were developed and refined as formulas that could be computed independently, using simple linguistic features from the texts themselves [54,55]. These readability indexes are easy to implement programmatically. Accordingly, they are embedded in popular text processing software.³

3. Proposal

Following the literature reviewed in Section 2, we identify a research gap that this work aims to address: the design, implementation, and validation of a combination of explainable-by-design FRBSs and CoT-enhanced LLMs that integrates factual and counterfactual explanatory narratives. The most similar work in the literature is the publication by Fredes et al. [45], but it lacks the flexibility to choose between factual and counterfactual insights and the trustworthiness provided by our underlying fuzzy explainer, as their proposal involves the LLM autonomously inferring the explanatory rules from counterfactual examples.

In this work, we address the following research hypotheses:

- H1: “The addition of counterfactuals to automatic factual explanations about a system’s predictions results in better explanatory quality through an increase in completeness that does not hamper reliability or trustworthiness.”
- H2: “CoT Prompting applied to discourse structuring and content planning helps LLM-based explainers produce clearer, more easily understandable narratives.”

3.1. Architecture overview

Fig. 1 illustrates the proposed pipeline to generate narrative explanations from tabular data. Initially, in the offline stage, a decision tree is built for each given dataset. Then, we apply linguistic approximation to translate the tree into an FRBS. Afterwards, in the online stage, given a single data instance, each FRBS makes a prediction that is justified with factual and counterfactual pieces of information. Then, the fuzzy explainer is able to fuse the information provided by multiple FRBSs into a single narrative explanation. Finally, an LLM rewrites such an explanation to get better naturalness without jeopardizing reliability and truthfulness.

Each stage in the pipeline will be discussed more extensively in the following sections.

3.1.1. Offline stage

The use case taken for demonstrative purposes focuses on medical diagnostics for diabetes and heart disease. Given the key importance of trust in the medical field, this use case is considered representative of areas where explainability, as a component of trustworthiness, is to be maximized. The datasets are as follows:

³ For example, the NLTK toolkit (<https://www.nltk.org/>) is a widely used Python library for NLP tasks and contains a readability module in its contrib version package. The popular Microsoft Word also natively supports the calculation of readability metrics.

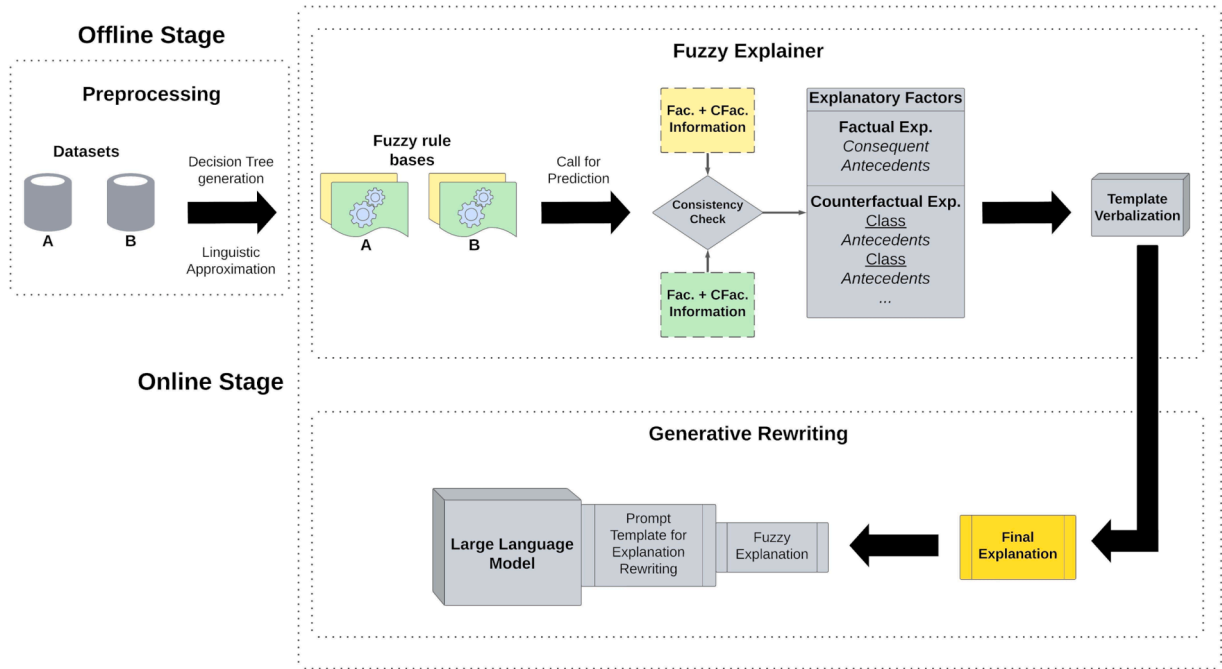


Fig. 1. Schematic view of the explainer pipeline followed, showcasing the offline stage in which explainable knowledge bases are created, and the online stage in which they are exploited to predict and explain for test instances through a combination of fuzzy inference, information fusion and LLM-based verbalization.

- The heart disease⁴ dataset comes from the 2022 edition of the annual CDC⁵ Behavioral Risk Factor Surveillance System (BRFSS), a telephone survey on the health status of American residents.
- The diabetes⁶ dataset comes from the 2015 edition of the same survey.

The two datasets are interlinked by virtue of them sharing their general feature space, as they are originated in the same survey (although in different editions). The shared feature space enables automatic cross-dataset checks that supplement the explanatory narratives with richer insights for the end user. For instance, in a counterfactual setting, the system could conclude that the risk of heart disease would be lower if body mass index (BMI) decreased, and also assess the side effects of such a change for diabetes, enriching the final utterance.

The original datasets are pre-processed so that only lifestyle and demographic factors are included in the generated fuzzy rule bases, in order to avoid more complex factors that could confuse non-expert users. After pre-processing, both datasets share the same set of attributes, with diabetes and heart disease swapping places as the dependent variable depending on the case. The attributes in the datasets are:

- **Gender**, binary (Male / Female).
- **Age**, binned in age groups described by a representative centroid age.
- **BMI**, numeric.
- **Smoker**, binary (No/Yes).
- **Physical Activity**, binary (No/Yes).
- **Alcohol Drinking**, binary (No/Yes) and dependent on heavy consumption (more than a certain amount per week).
- **Diabetic**, binary (No/Yes).
- **Heart Disease**, binary (No/Yes).

It is worth noting that, as medical datasets, some attributes should be considered protected variables from the perspective of AI fairness and ethics, most importantly, age and gender. This is largely unavoidable in the healthcare context, at least when predictive tasks are undertaken, as most relevant information for medical diagnosis is sensitive. As data instances are anonymized and do not hold enough information to identify the subjects being described, and because their use in this work is only demonstrative, we have not extensively addressed this concern, but it is still important to point out that, were these data sources to be used in a real-world application, careful management of those attributes would be warranted.

⁴ Personal Key Indicators of Heart Disease Dataset: <https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease>

⁵ CDC stands for U.S. Centers for Disease Control and Prevention, <https://www.cdc.gov>

⁶ Diabetes Health Indicators Dataset: <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>

Table 1

Summary of quality metrics for the base trees and random forest (performance on 10-fold cross-validation).

<i>Model</i>	Accuracy	Precision	Recall	F1-Score
Heart - RF	0.739	0.739	0.739	0.738
Heart - J48	0.730	0.734	0.730	0.729
Heart - REP	0.728	0.732	0.728	0.726
Heart - FDT	0.723	0.729	0.723	0.721
Heart - FDTP	0.701	0.701	0.701	0.701
Diabetes - RF	0.737	0.739	0.737	0.737
Diabetes - J48	0.701	0.701	0.701	0.701
Diabetes - REP	0.700	0.701	0.700	0.700
Diabetes - FDT	0.696	0.698	0.696	0.695
Diabetes - FDTP	0.674	0.674	0.674	0.674

Table 2

Summary of quality metrics for the base trees and random forest (predictive performance on the test datasets).

<i>Model</i>	Accuracy	Precision	Recall	F1-Score
Heart - RF	0.742	0.743	0.742	0.742
Heart - J48	0.721	0.726	0.721	0.719
Heart - REP	0.717	0.721	0.717	0.715
Heart - FDT	0.714	0.677	0.817	0.74
Heart - FDTP	0.696	0.699	0.687	0.693
Diabetes - RF	0.735	0.737	0.735	0.735
Diabetes - J48	0.689	0.689	0.689	0.689
Diabetes - REP	0.688	0.689	0.688	0.688
Diabetes - FDT	0.683	0.670	0.722	0.695
Diabetes - FDTP	0.671	0.670	0.674	0.672

Another important aspect of the attribute set is its actionability, meaning the extent to which its features can be changed. While BMI, smoking, physical activity, and alcohol drinking are actionable, biological gender and age clearly are not, and diabetes and heart disease status are also hardly possible to alter. Ideally, when aiming for prescriptive counterfactuals (i.e., counterfactual explanations that point out to a user how they could flip a decision), all or most attributes presented should be actionable. A descriptive counterfactual is not subject to this, as it aims to give the broadest perspective possible on the problem at hand instead of providing an avenue towards changing the final verdict. For our purposes in this work, we align with descriptive-style counterfactuals, which means that all attributes, and not only those that are actionable, will be allowed to be part of the counterfactual explanations.

As a final remark on dataset preprocessing, balance is enforced through subsampling (since the datasets are initially heavily unbalanced towards negative diagnoses, as is common in the healthcare domain), and both datasets are split 90/10 into training and test sets, resulting in final splits of 63622/7070 for diabetes and 49271/5475 for heart disease.

After these preparations, the training splits for each dataset are used to train Decision Trees [56], employing different algorithms to ensure diversity in the trees' decision paths. The generation of crisp trees was carried out by the J48 and REPTree algorithms implemented by the WEKA environment⁷. The generation of fuzzy trees was carried out by the FDT algorithm implemented by the GUAJE environment⁸, with FDTP being the pruned tree. Default parameters were used to generate both crisp and fuzzy trees. The training sets are also employed to fit a Random Forest (RF) [57], which serves as a non-interpretable "judge" intended to break ties if the decision trees do not agree on their output class when looking at new unknown cases.

The performance metrics of the generated tree models are reported in Tables 1 and 2.

Fuzzy trees generated by GUAJE, both FDT and FDTP, are endowed with global semantics by imposing the use of strong fuzzy partitions a priori [58]. As a result, each tree is stored as an FRBS. However, the crisp trees are subjected to a process of linguistic approximation, as described in [28,59], aided by the ExpliClas [60] online tool⁹. This transforms each tree into an FRBS, which changes the performance of the trees. Table 3 reports the predictive metrics of the fuzzified trees, along with the natively fuzzy trees, in both test datasets.

Because the fusion strategy of the fuzzy explainer is designed to work with two FRBSs, it is necessary to make a choice in which two of the four trees to utilize. Considering that the purposes of this work are aligned with explanatory quality rather than predictive performance, it is not adequate to just select the trees based on the metrics available above. Interpretability and informativeness are also important properties to consider, and in order to have a better outlook on how the models compare in those terms, while still considering accuracy-related metrics, Fig. 2 shows several Pareto fronts.

⁷ <https://ml.cms.waikato.ac.nz/weka/>

⁸ <https://gitlab.citius.gal/jose.alonso/guaje/>

⁹ <https://demos.citius.usc.es/ExpliClas/>

Table 3

Summary of quality metrics for the fuzzified (fuzz.) and native fuzzy trees (predictive performance on the test datasets).

Model	Accuracy	Precision	Recall	F1-Score	Rules
Heart - Fuzz. J48	0.702	0.687	0.740	0.713	52
Heart - Fuzz. REP	0.699	0.711	0.672	0.691	60
Heart - FDT	0.714	0.677	0.817	0.74	482
Heart - FDTP	0.696	0.699	0.687	0.693	6
Diabetes - Fuzz. J48	0.641	0.650	0.611	0.630	57
Diabetes - Fuzz. REP	0.660	0.722	0.521	0.605	72
Diabetes - FDT	0.683	0.670	0.722	0.695	549
Diabetes - FDTP	0.671	0.670	0.674	0.672	14

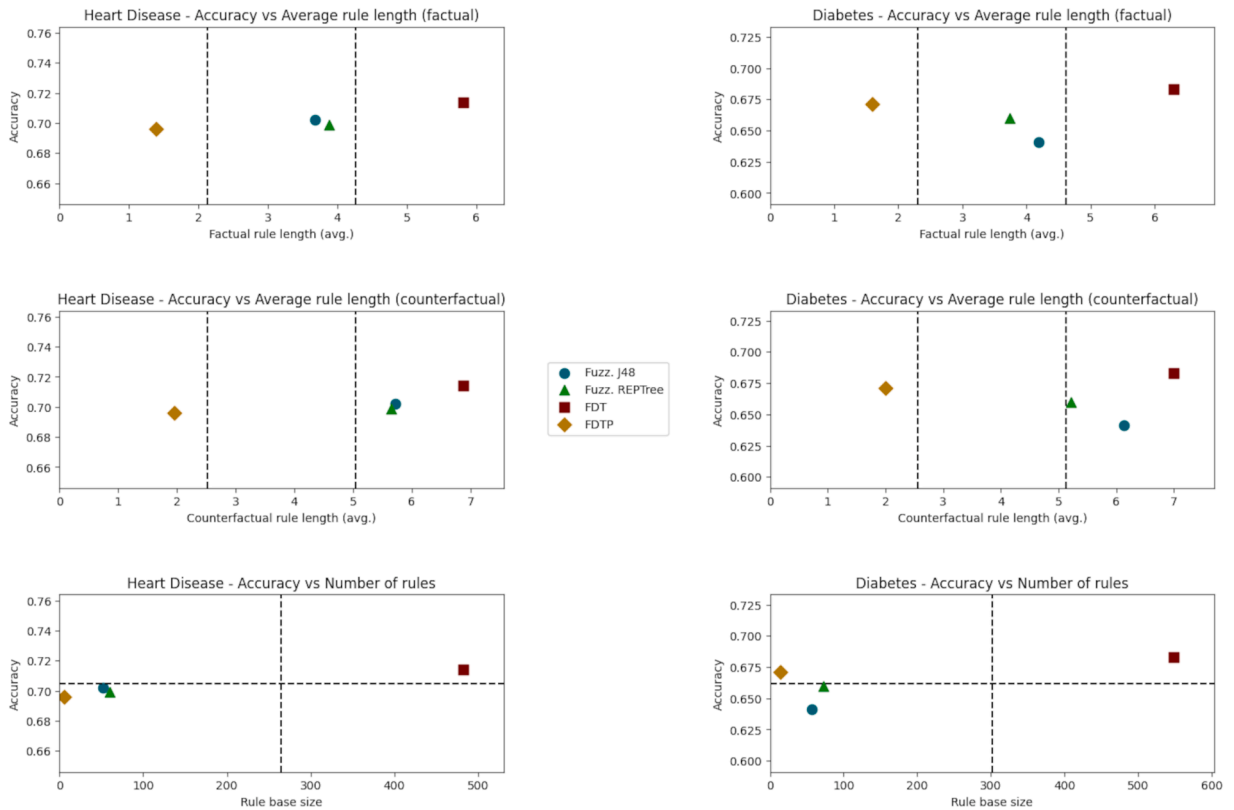


Fig. 2. Pareto fronts contrasting accuracy vs. interpretability for the test datasets (heart disease on the left column, diabetes on the right column) with average rule length regarding fired rules in factual cases (in the first row), average rule length regarding fired rules in counterfactual cases (in the second row), and number of rules (in the third row).

Regarding firing rules for computing factual (counterfactual) explanations, the average rule length (i.e., the number of premises divided by the number of rules) is taken as a surrogate metric for the informativeness of the explanations, while rule base size (i.e., the number of rules) is taken as a surrogate metric for the interpretability of the FRBS itself. Intuitively, better predictive performance along with smaller rule bases and/or shorter rules is preferable. This heuristic would take the FDTP as the best option, as it has comparable predictive performance while having very short rules and extremely small rule bases. The second-best option is harder to select, as both fuzzified crisp trees are very similar in all fronts.

However, upon qualitative evaluation of the FDTPs, it was found that the pruning process was overly aggressive, leaving only two features (namely, age and diabetes status for heart disease, age and BMI for diabetes) in each FRBS, resulting in exceedingly small rule lengths. This was deemed insufficient to produce useful explanations, as explaining every diabetes case solely by age and BMI does not appear to characterize the problem adequately. Accordingly, we discarded FDTPs as an option.

This type of qualitative analysis was repeated on the remaining trees, prompted by the previous findings. The full FDTs presented the opposite problem to their pruned counterparts: they consistently used too many features in their explanations, many times citing every single feature of the data instance as a factor. This is also undesirable, as explanations are to be presented alongside a description

of the subject, and citing every feature as part of the explanation is redundant with the description itself, since nothing is highlighted as a deciding factor behind the diagnosis.

Consequently, we opted for the fuzzified versions of the J48 and REP trees as the final pair of FRBSs in the fuzzy explainer, as they occupied a middle ground between providing too much and too little information, while staying comparable to natively fuzzy alternatives in terms of their predictive performance. Then, factual and counterfactual pieces of information extracted from each fuzzy system can be integrated into an explanatory interface that provides template-based natural language narratives, derived from a fusion of each rule base's inference for a given test instance, as detailed in the next section. The interface can also verbalize the data associated with test instances, including the subject descriptions mentioned above.

3.1.2. Online stage

When a prediction is issued, the consistency of factual outputs is checked, and in case of disagreement, the tie is broken by the Random Forest. This is a reasonably rare occurrence: it happens for 13.2% of the heart disease instances, and 18.3% of the diabetes subjects. Then, each fuzzy rule base independently generates factual and counterfactual pieces of information to justify the prediction. The aggregation of information is done by following the techniques detailed in Section 3.2. Finally, the aggregated information about the factual prediction and any potential counterfactuals is sent to a verbalization module, which uses templates to generate a natural language narrative. The resultant explanatory text is sent to the LLM-based component of the system, being embedded into a prompt that instructs the language model to rewrite the information received into a more natural, human-like form, without altering its key elements (so as not to lose the truthfulness and trustworthiness that operating with a fully explainable, data-grounded rule base provides). The generated narrative is taken as the final explanation to be presented to a potential user. The different prompting alternatives employed are detailed in Section 3.3.

3.2. Fusion of factual and counterfactual pieces of information

As discussed in Section 3.1, a key part of the proposed architecture is the fusion between explanations obtained from similar, yet diverse fuzzy rule bases. In implementation terms, this fusion relies on the features and values that are included in each fired rule (both for factual and counterfactual cases). However, considering that those features and values are, in reality, the linguistic variables and terms coming from each FRBS, it is proper to formally introduce the corresponding notation before describing the fusion strategies themselves.

3.2.1. Preliminary notions

Adhering to the notation given in [11], we introduce a fuzzy set A as was originally defined by Zadeh [7], by identifying a membership function that maps A from an Universe of Discourse U to a completely distributive lattice in the interval $[0, 1]$:

$$\mu_A : U \mapsto [0, 1]$$

with $U = [m_U, M_U] \subset \mathbb{R}$ being a one-dimensional subset of the real line with lower bound m_U and upper bound M_U .

It is worth noting that we will use fuzzy sets to characterize linguistic propositions like those considered in a Mamdani-type FRBS [61]. An FRBS is made up of a Knowledge Base (KB), which includes linguistic variables (LVs) and fuzzy IF-THEN rules, and a fuzzy inference processing structure capable of exploiting the rules. According to Zadeh [62], an LV is defined as a tuple $\langle X, T, U, G, \mu \rangle$ where:

- X is the name of the variable that is being modeled.
- T is a set of linguistic terms, i.e., symbols most normally expressed in natural language. Each term is characterized by a fuzzy set A .
- U is the Universe of Discourse that is associated with the domain of applicability for X .
- G is a grammar capable of generating the symbols included in T .
- $\mu : T \mapsto \mathcal{F}(U)$ corresponds to the semantic interpretation of the symbols in T . The interpretation of each term is characterized by a membership function μ_A for a fuzzy set A in the fuzzy partition $\mathcal{F}(U)$. In addition, in the case of strong fuzzy partitions, it holds that $\sum \mu_A = 1$.

For example, Age is an LV that may be described with linguistic terms such as Young, Adult, or Senior in the interval $[0, 100]$. Each term describes how much a certain numerical age would “fit” a symbolic description. Thus, $\mu_{Senior}(30) = 0$ means that a person P1 with 30 years would never be described as a senior person. Similarly, $\mu_{Adult}(30) = 1$ means the same person would always be described as an adult person. Moreover, considering a strong fuzzy partition for Age, if $\mu_{Senior}(27) = 0.7$ then $\mu_{Adult}(27) = 0.3$ and this means the second person P2 would be described as an adult person too, but with a smaller membership degree than P1.

Following Zadeh's definition of a linguistic proposition as an atomic *soft constraint* [63], given a value x in U , we interpret the degree of truth of X is A (e.g., Age is Young) regarding the membership degree $\mu_A(x)$. This can be extended to more complex constraints, in which several atomic constraints are linked by logical operations such as conjunction (AND) or disjunction (OR), to be implemented by t-norms and s-norms, respectively. For the purposes of this work, conjunctive soft constraints are considered, as they naturally follow from the initial Decision Tree structure, in which the branch leading to a given leaf node is a conjunction of many individual conditions:

$$X_1 \text{ is } A_{X_1} \text{ and } X_2 \text{ is } A_{X_2} \text{ and ...and } X_n \text{ is } A_{X_n}$$

where A_{X_i} represents the linguistic term associated to the corresponding variable X_i in each soft constraint.

The same kind of constraints are used to define the premises of IF-THEN rules in a Mamdani-type FRBS as follows:

$$R^k : \text{IF } X_1 \text{ is } A_{X_1} \text{ AND } X_2 \text{ is } A_{X_2} \text{ AND } \dots \text{AND } X_n \text{ is } A_{X_n} \text{ THEN } Y \text{ is } A_Y$$

where the logical operators AND and THEN are implemented as the minimum t-norm.

A simplified representation of the rule R^k is as follows:

$$R^k : \text{IF } p^k \text{ THEN } c^k$$

$$R^k : \text{IF } p_1^k \text{ AND } \dots \text{AND } p_n^k \text{ THEN } c^k$$

where the conclusion c^k of the rule is the claim to be justified regarding its premise, which has a conjunctive form with n soft constraints named in short as $p_j^k = (X_j \text{ is } A_{X_j})$.

Our fuzzy explainer translates the branches in a given decision tree into the IF-THEN rules of a Mamdani-type FRBS. Moreover, the crisp numerical constraints in the nodes of the tree are translated into atomic soft constraints that are verbalized in natural language, thanks to the mapping between fuzzy sets and linguistic terms that is naturally done with LVs. For example, if given a data instance, the winner rule, i.e., the rule with the highest firing degree in the FRBS were " WR^1 : IF Diabetic is No and Age is Adult THEN HeartDisease is No", the linguistic verbalization of such a rule (e.g., "The person has no heart disease because is a non-diabetic adult") would be taken as the factual explanation of the inferred claim (i.e., the consequent of the winner rule: "The person has no heart disease") for the observed case. On the other hand, when seeking a counterfactual explanation, we first need to identify the closest rule that would become the winner rule if the constraints for the alternative class were met (e.g., " WR^2 : IF Diabetic is Yes and Age is Adult THEN HeartDisease is Yes"). Accordingly, the linguistic verbalization of such a rule with the focus on the distinct constraints in WR^2 while omitting the shared constraints with respect to WR^1 (e.g., "The person would have heart disease if she were diabetic") would be taken as the counterfactual explanation of the observed case.

3.2.2. Fusion strategies

Given a set of LVs that establishes the global semantics for the problem under study, we can fuse as many fuzzy rule bases as there are available decision trees. In practice, we have a single Mamdani-type FRBS with fuzzy rules coming from different trees as if they were pieces of knowledge provided by different physicians. Fortunately, we can compare all the rules at the linguistic level to identify potential inconsistencies and/or redundancies because they all share the same definitions of the LVs. Without loss of generality, for the rest of this section, we will assume that we have only two trees (i.e., DT^1 and DT^2) that are translated into two sets of rules (i.e., RB^1 and RB^2) ready to act as two fuzzy explainers (i.e., EXP^1 and EXP^2) whose factual and counterfactual explanations are to be merged.

On the one hand, there is no fusion when the factual explanations given by EXP^1 and EXP^2 do not agree on their claims. In this case, the tie is broken by an auxiliary oracle Random Forest, and the explanations (both factual and counterfactual) associated with the winner explainer (i.e., the explainer in agreement with the oracle) are verbalized afterwards.

On the other hand, in case there is an agreement on the claims posed by EXP^1 and EXP^2 , then we can proceed with the fusion of factual explanations. We pay attention to the winner rule for each explainer (i.e., WR^1 and WR^2). Rule consequents themselves become irrelevant (other than to enunciate them on the final explanation), and thus the explanatory interface operates exclusively on the premises p^1 and p^2 of rules WR^1 and WR^2 , respectively:

$$p^1 = p_1^1 \text{ AND } \dots \text{AND } p_n^1, \text{ with } p_i^1 = (X_i \text{ is } A_{X_i}), 1 \leq i \leq n, \text{ and } 1 \leq n \leq N$$

$$p^2 = p_1^2 \text{ AND } \dots \text{AND } p_m^2, \text{ with } p_j^2 = (X_j \text{ is } A_{X_j}), 1 \leq j \leq m, \text{ and } 1 \leq m \leq N$$

where each premise can involve a different number of soft constraints (i.e., in the general case, it can be $n \neq m$), and N is the number of LVs. The superscripts (1 or 2) refer to the index of the corresponding rules.

The fusion maxim is to utter first the common explanatory factors (i.e., the shared premises such as $p_i^1 = p_j^2$, which means the same LV is described with the same linguistic term in both premises), and to add later other explanatory factors exclusive to one of the explanations as additional support to the inferred claim. In this manner, higher importance is placed on those soft constraints that are considered relevant by both explainers.

The set of unique explanatory factors for EXP^1 is as follows:

$$UF^1 = \{p_i \mid p_i \in p^1; p_i \notin p^2\} \quad (1)$$

The set of unique explanatory factors for EXP^2 is as follows:

$$UF^2 = \{p_i \mid p_i \notin p^1; p_i \in p^2\} \quad (2)$$

The set of shared explanatory factors is as follows:

$$SF = \{p_i \mid p_i \in p^1; p_i \in p^2\} \quad (3)$$

The set of additional joint supplementary explanatory factors is as follows:

$$AF = \{(p_i, p_j) \mid p_i \in p^1; p_j \in p^2; X_i = X_j; A_{X_i} \neq A_{X_j}\} \quad (4)$$

AF includes premises associated with LVs that appear in both explanations, but with different linguistic terms assigned. This means that both explanations have a soft constraint over the same variable, but they are not related to the same linguistic term. This may occur when membership functions associated with different linguistic terms overlap, as is typically the case with adjacent terms in strong fuzzy partitions. Even if it is not a mistake, this situation must be properly handled to prevent providing final users with ambiguous and misleading information. In order to do so, the linguistic term that will be included in the verbalized explanation is the one with the higher membership degree. Then, the verbalization process is applied as follows:

- If both SF and AF are empty, then there are necessarily unique explanatory factors to be considered. Otherwise, there would be no information to elaborate on a plausible explanation. If this is the case, then UF^1 and UF^2 are considered the primary explanatory factors, and they are listed sequentially.
- In case either SF or AF are non-empty, they are verbalized as the main factors in the explanation. Then, if either UF^1 or UF^2 or both are non-empty too, they are sequentially introduced as supporting evidence. Otherwise, supporting evidence can not be given.

Regarding the generation of counterfactual explanations, the approach is similar to the aforementioned process for generating and verbalizing factual explanations. For each set of rules, we search for the rule that would be the winner for producing the desired claim with a minimal change with respect to WR^1 and WR^2 , i.e., WR^{c2} and WR^{c1} , respectively. It is worth noting that there is no need for consequent agreement in the counterfactual mode, as the system is no longer providing a prediction grounded in the data of an instance, but crafting a hypothetical scenario in which the consequent would be different from that of the factual context. The new sets of explanatory factors are as follows:

$$UF^{c1} = \{p_i \mid p_i \in p^{c1} \text{ and } (p_i \in AF^c \text{ or } p_i \notin p^{c2})\} \quad (5)$$

$$UF^{c2} = \{p_i \mid p_i \in p^{c2} \text{ and } (p_i \in AF^c \text{ or } p_i \notin p^{c1})\} \quad (6)$$

$$SF^c = \{p_i \mid p_i \in p^{c1}; p_i \in p^{c2}\} \quad (7)$$

$$AF^c = \{(p_i, p_j) \mid p_i \in p^{c1}; p_j \in p^{c2}; X_i = X_j; A_{X_i} \neq A_{X_j}\} \quad (8)$$

If none of the explainers can produce the desired counterfactual, then no counterfactual explanation is offered to the user. If only one explainer can not produce a counterfactual for the given output class, then the other explainer is utilized (computing either UF^{c1} or UF^{c2} for the counterfactual winner rule; see Eqs. (5) and (6) respectively). Otherwise, SF^c (see Eq. (7)) and AF^c (see Eq. (8)) are computed. Once again, the priority is placed on offering the shared factors of the counterfactual rules as the main elements leading to the change in the predicted claim. However, in this situation, it is not necessary to disambiguate AF . In a factual context, the disagreement in linguistic terms means a different restriction over the same underlying value (the value on the actual data instance being classified). As the inference engine generates counterfactuals by seeking the smallest perturbations to the data instance that would lead to a change in classification, in the counterfactual context this underlying value need not be shared by the two winner rules; this means that their disagreement in linguistic terms is not an ambiguous interpretation of the same value, but different hypotheticals that could change the consequent.

Accordingly, the elements in AF^c , when verbalizing counterfactuals, are given the same treatment as elements in either UF^{c1} or UF^{c2} , i.e., as supporting evidence for the elements in SF^c . Semantically, each counterfactual winner rule constructs a hypothetical scenario that leads to a different outcome; some of the “hypotheses” may be shared, and thus given more importance, but those that are not shared are equally valuable alternatives to reach the counterfactual outcome. It is worth noting that those non-shared factors cannot, however, be offered sequentially as was done before. On the one hand, in factual explanations, all the constraints are equally valid because the data instance serves as common ground for the fired rules, and so constraints can be mixed together once disambiguation is performed. On the other hand, in counterfactual explanations, constraints represent information extracted from different scenarios, and unless they are shared (which means a point of agreement), they can not be mixed.

The verbalization of counterfactual explanations is as follows:

- If UF^{c1} , UF^{c2} , and SF^c are non-empty sets, then the shared factors (i.e., SF^c) are verbalized first. Then, it is also stated that those factors would need to be complemented by one of two additional sets of constraints, which are the explanatory factors in UF^{c1} and UF^{c2} .
- If SF^c is a non-empty set of explanatory factors, but only one of the two unique explanatory factors, either UF^{c1} or UF^{c2} is a non-empty set, then we apply the same verbalization process as in the factual case. Thus, shared factors are presented as the core of the explanation, with the unique factors serving as supporting evidence. This is done because in this situation, one of the explanations is strictly more specific (i.e., one of the counterfactual winner rules imposes further constraints on a situation that the other would already consider sufficient for a change in the consequent).
- If SF^c is a non-empty set, but both UF^{c1} and UF^{c2} are empty sets, then it means that both hypothetical scenarios were fully equivalent, and it is only necessary to verbalize the shared factors.
- If SF^c is an empty set, then both UF^{c1} and UF^{c2} are non-empty sets because otherwise, one of the counterfactual rules, or both, would not have produced a counterfactual and no fusion would be done at all. In this case, both sets of unique explanatory factors (i.e., UF^{c1} and UF^{c2}) are verbalized separated by a disjunction, presenting them as equally valid alternatives.

It is also important to note that the implemented explanatory interface allows for two methods of information extraction in the counterfactual case, as it can either include the constraints already present in the sets pertaining to the factual case or omit them. This is because when verbalizing counterfactuals as stand-alone explanations, all explanatory factors are of interest; but when using them to supplement an initial factual explanation, it is generally undesirable to repeat constraints that were already imposed and verbalized as part of the factual piece of information. For example, one may consider a situation in which the factual explanation for *HeartDisease is Yes* includes the constraints *Age is Senior* and *Diabetes is Yes*. If the counterfactual explanation for *HeartDisease is No* includes the constraints *Age is Senior*, *Diabetes is No* and *Smoking is No*, a possible verbalization would be “The person presents heart disease because they are a senior and they present diabetes. They would not present heart disease if they were a senior, they did not present diabetes and did not smoke”. Because the factual explanation was enunciated first, the repetition of a constraint is useless and can be misleading.

To complement this descriptive analysis of the fusion strategies, A contains an end-to-end example of how this procedure is followed to create a textual narrative from the fired rules.

3.3. Prompting approaches

The addition of CoT to the system is not difficult, as the greatest draw of the prompting paradigm is that complex and highly beneficial behavior can be elicited through a natural language instruction. The stochastic nature of LLMs, however, makes optimizing prompts non-trivial, and thus it is relevant to discuss the prompting structure that we applied.

Most CoT prompting schemes are included to enhance the resulting system’s skill on a task. In this work, the task being improved is that of rewriting an existing explanatory narrative, without interfering in the prediction step, so as not to lose the truthfulness and trustworthiness of information coming from the FRBS. In this vein, our CoT prompt is directed towards content planning and discourse structuring [64,65].

Explanations that are either purely factual or a mix of factual and counterfactual pieces of information can utilize the same prompt. This is because the prompt instructs the model to rewrite the explanation: a factual explanation serves as the baseline case, while the contrast between the factual and counterfactual segments implicitly highlights the difference between the two in a mixed explanation.

In case that a purely counterfactual explanation is required, i. e., an explanation that does not contain factual information at all, it is convenient that the prompting scheme makes this fact clear to the LLM because in absence of a factual explanation against which to be contrasted, counterfactuals are confusing (see Section 2.1 for a more in-depth discussion of the contrastive nature of counterfactuals).

The general structure of the prompts is as follows:

- The initial sentence gives context of the role the LLM is supposed to play: “for this use case, an expert assistant helping users learn about an illness”.
- Then, the data that will be provided to the LLM (i.e., the test subject’s description and explanations) is described. In the purely counterfactual case, this style of explanation is also detailed.
- Next to its description, the data is directly embedded in the textual form produced by the fuzzy component.
- After the embedded data, the prompt describes the task to be performed: “to rewrite the explanation portion into a coherent, more natural explanatory narrative”. Several instructions are added to ensure that the model does not alter existing facts or add new details. In the exclusively counterfactual case, the task description is slightly adapted to better fit the semantics of a counterfactual.
- Once the task has been described, the CoT prompt is added: “instructing the model to not write anything right away, and to always perform a full reasoning about what it must write in order to comply with the full set of instructions given”. Notice that the type of CoT being utilized is of the ‘zero-shot’ variant, as no examples of successful task completion are given to supplement the CoT instruction itself.
- Finally, some specific formatting details are added. In case the system was to be used as a chat model, no specific format is needed, but for experimental settings it is convenient to instruct the model to produce more structured outputs.

For reference, Appendix B details all the prompt templates that were designed and utilized during the experimental validation of this work.

3.4. Validation metrics

In order to validate the research hypotheses, and to measure the effectiveness of the two contributions of this work (i.e., the new counterfactual modality and the addition of CoT prompting for text structuring), it is necessary to rigorously evaluate the final explanatory narratives. While this would be best done through a human study, as works on explainable AI performance measurement such as [66] posit, its design and execution are complex enough that performing one was considered beyond the scope of the current work, and it was kept as a future effort. Automatic evaluation was chosen instead, to serve as an initial gauge of the validity of the proposed contributions, even though we acknowledge the limitations of such a choice.

For this evaluation, a total of seven metrics have been chosen to act as surrogates for trustworthiness and explanatory quality:

- **Predictive performance:** we must verify that the rewriting step performed by the LLM does not compromise the system’s performance in terms of prediction, compared to the underlying fuzzy explainer. It is also needed to compare the performance of our approach against purely generative architectures, where fuzzy information is not supplied. To obtain it, we ask the LLM to

give its judgment on the output class according to the given explanation, along with its rewriting. Considering that validation is being performed in the medical context, where false positives and negatives may have different relevancy, four metrics related to prediction will be computed:

- **Accuracy:** computed as the proportion of correct predictions, regardless of their class. The more general measurement, it is appropriate for this work only due to the preprocessing balancing the datasets beforehand.
- **Precision:** computed as the proportion of subjects where a positive prediction was made, that actually were positive. Lower scores in this metric means the approach is giving out too many positive predictions.
- **Recall:** computed as the proportions of actually positive instances that were classified as such. The higher this metric is, the higher share of positive cases that the approach finds.
- **F1-Score:** computed as the harmonic mean of precision and recall, it penalizes unbalanced scores in those metrics.
- **Response length:** computed as the length of utterances. We take this metric as a contributor to overall explanatory quality, as user studies on explainability have shown that proper brevity and simplicity are important for communication and effective explanation [48]. Generative models have also shown a preference for verbosity even when it is unnecessary [67], and the current contributions limiting this effect would be very positive.
- **Hallucinations:** computed as the number of attributes of the instance that show potentially hallucinated content. Hallucinations not only degrade the user satisfaction, but they also jeopardize trustworthiness [68], making this measurement necessary. The reduction of hallucinations (due to the addition of on-demand fuzzy-grounded information) was the main goal of our previous work [13]. For the purposes of this work, it is necessary to check that the new modalities and prompting strategies are not increasing hallucinatory tendencies, thus being in conflict with the advances of previous work.
- **Readability index:** computed in the form of the SMOG Index, which is considered the most suitable readability measurement in the healthcare domain [69], where it is crucial that medical materials are understandable to a wide audience. The most direct metric we'll employ to measure the quality of the explanatory narratives, SMOG describes the years of education that a person should have undergone in order to properly understand the evaluated piece of text, without the need for auxiliary transformations, and consequently acts as a measurement of how clear the final explanations is. As a brief note, while computing the SMOG Index usually requires the textual materials to be at least 30 sentences long, there are mathematical adaptations that make it suitable for shorter contexts. This work will rely on them, as explanatory utterances from the LLM will rarely meet the 30 sentence mark.

The accuracy and hallucination metrics primarily focus on trustworthiness, while response length and readability are mostly concerned with explanatory quality. It is worth mentioning that, in choosing these metrics, we were inspired by Grice's conversational maxims [70]. These maxims act as high-level principles that define fruitful communicative efforts, and as such, can help characterize what a good explanation entails. For example, the applicability of Gricean maxims to causal explanation was introduced by [71], considering explanation inherently conversational.

It is also worth noting that hallucinations are hard to identify automatically. Accordingly, there are several competitions (e.g., SHROOM¹⁰) with a focus on this automatic detection task. Among the most promising approaches, it is common to rely on the judgments of other LLMs during this task (e.g., HaluEval [72]). We consider chaining LLM calls to serve as judges to run contrary to our focus on trustworthiness, because an LLM-as-a-judge can also suffer from hallucinations. Consequently, a statistical measure has been chosen to measure hallucinatory content in the utterances of the final system. This measure was first introduced in [13], and it is based on a reverse matching of the generative responses of the LLM against the explanatory factors given by the fuzzy explainers. See Fig. 3 for an overview on how this matching is performed. In essence, a dictionary of regular expressions is computed for any given value that each of the attributes of the problem can take. The explanatory factors on the fuzzy explanation are taken to serve as the gold standard, and as such, need not be searched for (their presence is correct). Any other values detected through their regular expressions are considered to constitute a hallucination for their attribute, as their presence is not grounded in the fuzzy inference.

3.5. Experimental setup

Validating the new features of the explanatory interface requires rigorous comparative testing, to gauge potential performance gains derived from counterfactuals and CoT prompting. The basic explanation modality, which was available prior to the additions of this work, would be a purely factual explanatory style with no CoT. The addition of counterfactuals to the explanatory interface immediately spawns two new modalities: purely counterfactual, and mixed (where factual and counterfactual information are given as a combined whole). It is also convenient to have a baseline approach as a starting point for our metrics: a scenario where the LLM will not have access to the fuzzy explanatory interface, being asked to complete the prediction and explanation task based on the subject description alone. This makes for four different scenarios, in terms of the explanation modality.

Additionally, the usage of CoT prompting is a binary feature of the explanation generation process, and thus for each explanation modality, we may perform the task with or without CoT. That makes the number of scenarios grow to eight. When referring to these experimental scenarios, the modality will be mentioned first, with the presence or absence of CoT prompting being clarified after.

The full test splits that were set apart at the beginning of the system development were tested on these eight experimental scenarios, for both datasets: 7070 instances for diabetes, and 5475 instances for heart disease. Considering the data volume and the role of LLMs in producing the final narratives over which the validation metrics introduced in Section 3.4 were to be taken, the testing process was broken down in two distinct steps.

¹⁰ <https://github.com/Helsinki-NLP/shroom>

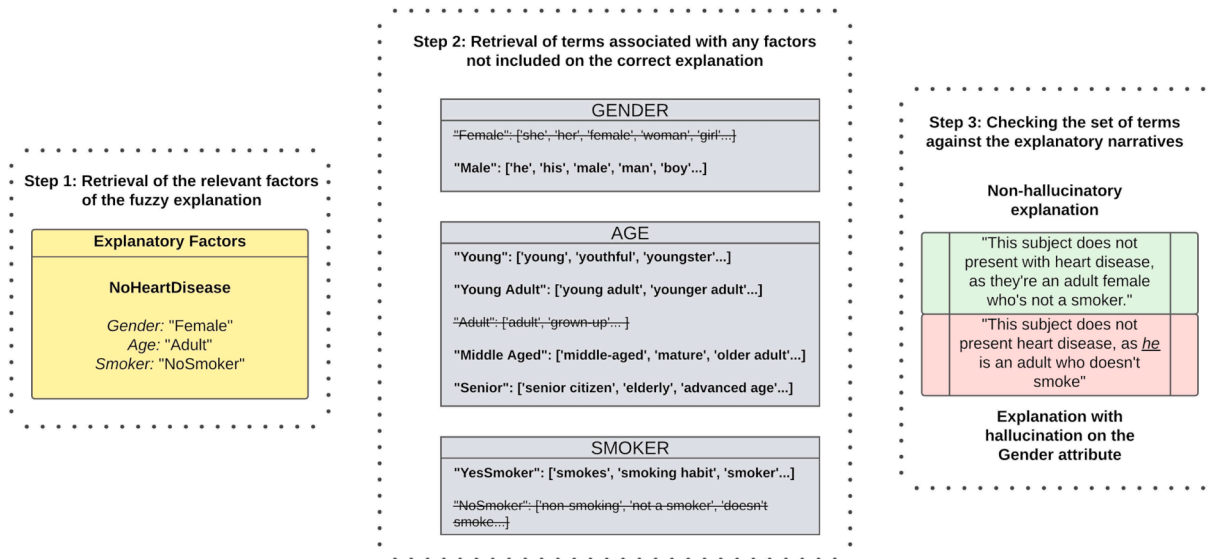


Fig. 3. Illustrative example for the hallucination detection method followed in this work, summarizing its key steps in detecting potential hallucinatory content in the explanations.

Firstly, a process of batch fuzzy inference ran the entire test sets through the fuzzy explainer. This involved predicting for the instances on both FRBSs, in factual and counterfactual forms, and merging the fired rules according to the strategies introduced in [Section 3.2.2](#) to then verbalize the resulting explanations through templates. All the intermediate results of this process are logged in the following JSON format:

- **ID**
- **True class**
- **Predicted class**
- **Fused**; describing if the Random Forest had to be used to break a disagreement between the trees.
- **Subject verbalization**; in which only the instance description is included.
- **Full explanation**; which merges factual and counterfactual information.
- **Factual explanation**
- **Counterfactual explanation**
- **Full counterfactual explanation**; as was introduced at the end of [Section 3.2.2](#), counterfactual can be presented with or without repetition of factors present on the factual counterpart, this field contains the with-repetition version of the counterfactual explanation.
- **Explanatory factors**; a nested field that holds the consequents and antecedents for both the factual and counterfactual explanations, along with their lengths in terms of factors, after merging the results from both trees.
- **Tree A results**; a nested field that holds the consequents and antecedents for both the factual and counterfactual explanations, along with activation strength of the consequent and membership degree of the antecedents, for Tree A.
- **Tree B results**; equivalent to the previous field for the remaining tree.

Notice that the system, when interacting with users, performs the call to the fuzzy explainer and the call to the LLM as a single process. However, this first pass of fuzzy inference allows us to optimize the computational load of the experiments, as the intermediate results can be reused for the LLM calls of all eight scenarios without repeatedly calling the fuzzy explainer.

Then, as the second step, each experimental scenario was implemented in such a way that the LLM sequentially rewrote the appropriate explanations (or generated them, for the baseline scenarios) for all instances, using the results of the previous step. To create the final result logs, also in JSON format, the utterances of the model were parsed and merged with some information from the intermediate results that was carried over. These final logs follow this format:

- **ID**
- **True class**
- **Fuzzy explainer class**
- **Fused**; with the same meaning as above.
- **LLM class**
- **Template subject verbalization**
- **LLM subject verbalization**

- **Template explanation;** which may be factual, counterfactual, or mixed depending on the scenario.
- **LLM explanation**
- **LLM reasoning;** only present on scenarios with CoT, contains the reasoning chain of the model.
- **Explanatory factors;** with the same meaning as above.

Notice that, on the baseline scenarios, the template explanation and explanatory factors are absent from the logs, as they are not utilized by the LLM. The template-based subject verbalization is still present because those scenarios do need a subject description to explain.

Once the result files were complete for all scenarios, validation metrics were computed using their information: predictive performance metrics are taken based on the “class” fields, while the remaining metrics (length, readability and hallucinatory content) are computed based on the textual narratives themselves. All the metrics described in Section 3.4 were computed for all scenarios, with the exception of hallucinations, which were only tested on the purely factual and purely counterfactual scenarios. They could not be tested in the baseline scenarios because the detection approach employed requires the explanatory factors provided by the fuzzy interface to serve as the gold standard. While the mixed case does have a set of fuzzy explanatory factors, the detection approach is too naive to work for that scenario, as a certain fragment may be taken or not as a hallucination depending on its placement when working with a mixed explanation.

As an illustrative example, consider the following explanation: “The person does not have heart disease because they are young and don’t have diabetes. The person would have heart disease if they were a senior.”. In case the explanatory factors included *Young* for the factual case and *Senior Citizen* for its counterfactual, the explanation is correct. However, if the cases were reversed, the explanation would be hallucinated. The current method would always ignore both particles, as they are both considered part of the set of explanatory factors, and thus would not be able to detect this subtler hallucination. This is a limitation of our hallucination detector to be addressed in future work.

The entire experimental process was conducted on a single PC, equipped with a 24-core Intel Core Ultra 9 285K CPU, 64GB of RAM and a 32GB Nvidia GeForce RTX 5090 GPU, running Ubuntu 24.04.4 LTS as the OS.

For the sake of open science, everything required for the reproducibility of the experiments is available in the following GitLab project: <https://gitlab.citius.gal/pabloomiguel.perez/fcflm-citius-ijar>

4. Experiments

In this section, we begin with an in-depth discussion of quantitative results (Section 4.1). Then, we proceed to a qualitative analysis of illustrative examples (Section 4.2).

4.1. Quantitative analysis of the reported results

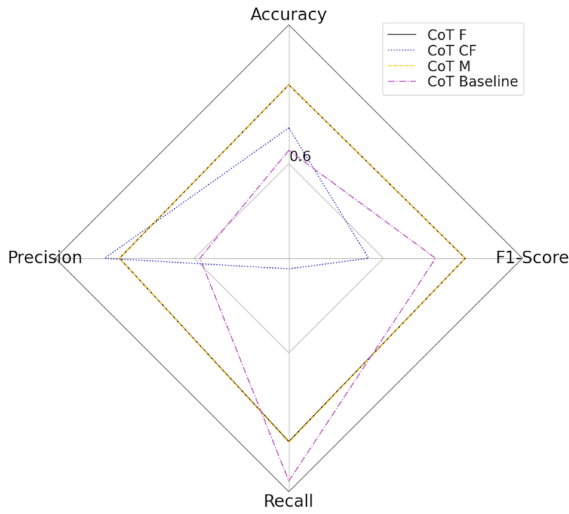
Figs. 4–7, alongside Table 4 report the experimental results obtained for the metrics described in Section 3.4. For completeness and ease of reference, Appendix C details all the numerical results in full, in tabular form.

To start with, the predictive performance of the different experimental scenarios is shown in Figs. 4 and 5, which showcase the results for the heart disease and diabetes datasets, respectively. Metrics are reported regarding: i) the Ground Truth (i.e., true labels of the instances); and (ii) the Fuzzy Inference (i.e., predicted class by the fuzzy explainer before the LLM rewrites the explanatory narrative). Notice that the baseline scenarios do not report their performance versus the Fuzzy Inference, as they do not use that information in their predictions. For clarity, CoT and Non-CoT scenarios are plotted separately in the figures.

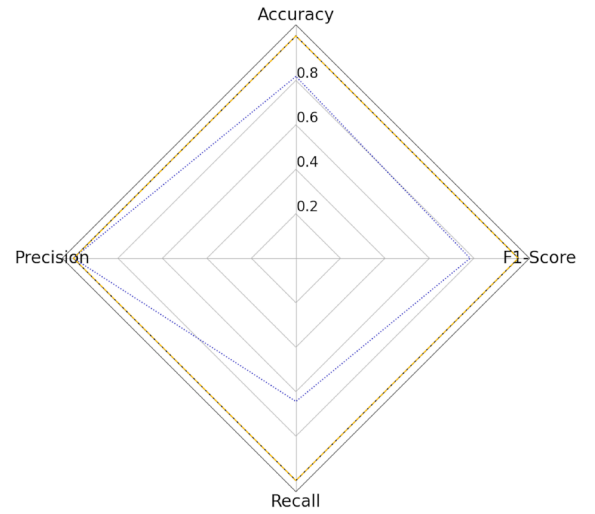
From the reported results, we can observe several trends. Looking at performance against ground truth, the behavior of the fuzzy-assisted scenarios is mostly stable between datasets: the factual and mixed cases are so similar in all the performance metrics that they mostly overlap in the radar plots, being hard to tell apart, and they are also the most stable approaches, with balanced metrics in all four dimensions which contribute to their accuracy and F1-scores being very similar and the best overall. The counterfactual case, however, seems much more conservative in assigning positive predictions, with good precision traded for poor recall, an imbalance that hurts the final F1-score to the point of ending up worse than the baseline in the heart disease tests. The baseline itself is the most unstable alternative: while it usually exhibits opposite behavior to the counterfactual case, being very prone to assigning positive predictions, for the diabetes dataset and with CoT, it flips into the same precision-first trend shown by the counterfactuals.

When dealing with performance against the fuzzy inference, the results are generally very good. The factual and mixed cases align very well with the fuzzy predictions, while the counterfactual approach deviates slightly from them, still showing a certain bias towards precision, most clearly against the ground truth. This deviation is most apparent in the heart disease experiments (see right-hand side in Fig. 4). These findings may indicate that the more straightforward factual and mixed scenarios capture and reproduce the FRBS’s decisions without difficulty, which is pivotal for trustworthiness as it proves strict adherence to the transparent system beneath the opaque LLM, while the purely counterfactual approach finds it harder to do so. This may be explained by the greater complexity of its task, as it must produce a factual prediction when only receiving counterfactual information from the fuzzy explainer.

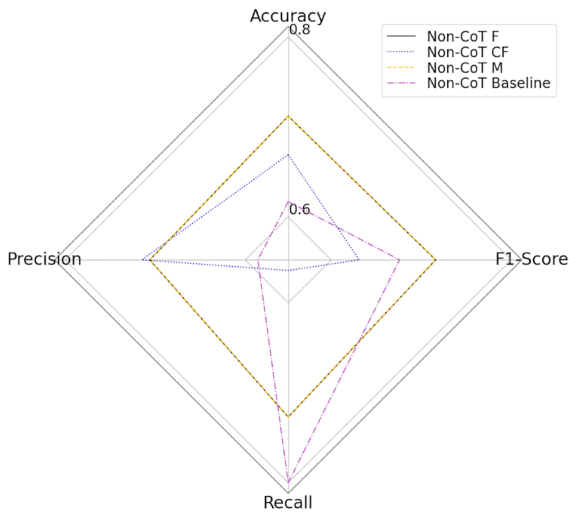
The effects of CoT seem negligible across the different experimental setups, outside of flipping the behavior of the baseline approach in the diabetes dataset (which was already noted in Fig. 5). Considering that the purpose of CoT prompting was to improve the explanatory quality of the final narratives, without affecting the predictive step at all, this is not a negative result for the fuzzy-assisted approaches, which are stable in prediction regardless of CoT status. It is reasonable that the baseline approach would be most vulnerable to the effects of CoT “leaking” outside their purported domain, as it does not have any additional information at its disposal to inform its final decision.



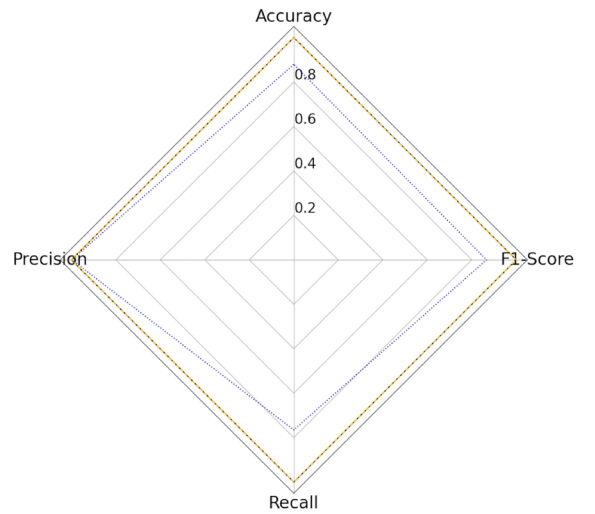
(a) Ground Truth - CoT



(b) Fuzzy Inference - CoT



(c) Ground Truth - Non-CoT

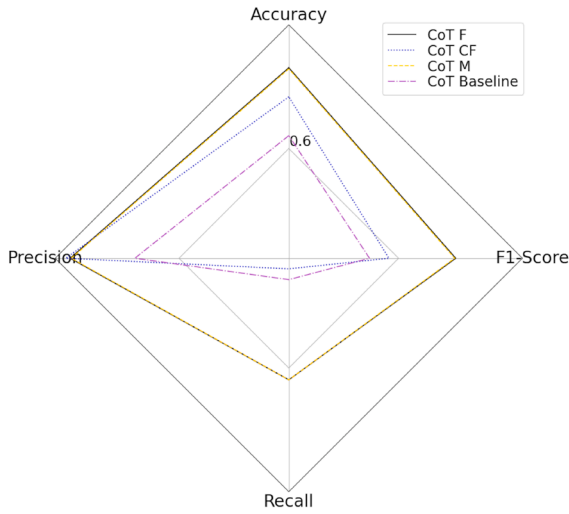


(d) Fuzzy Inference - Non-CoT

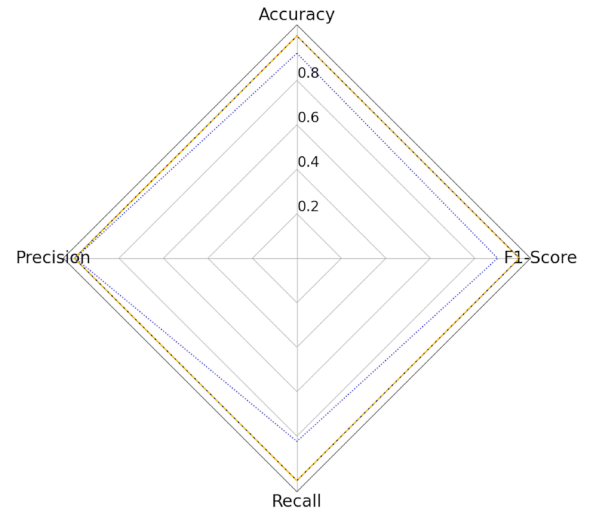
Fig. 4. Summary of the predictive performance achieved by the approaches during testing on the heart disease dataset. The first row, i.e., radar plots (a) and (b), shows the values of quality metrics in the scenarios with CoT. The second row, i.e., radar plots (c) and (d), shows the values of the quality metrics for the scenarios without CoT (i.e., Non-CoT). All metrics are reported against the ground truth (plots in the left: a and c) and against the fuzzy inference (plots in the right: b and d). The baseline approach lacks data for the latter.

Fig. 6 reports the average response length for the different scenarios. For each experimental setup, two bars are shown: the leftmost for the heart dataset and the rightmost for the diabetes dataset. Each bar is divided into two, with the lower bar reporting the length of the prediction explanation, and the upper bar reporting the length of the subject description (i.e., the verbalization of the test instance features). Because of how these metrics are computed, the total height of the bar, combining both explanation and subject description, is equivalent to the average length of the full explanatory narrative provided by the LLM.

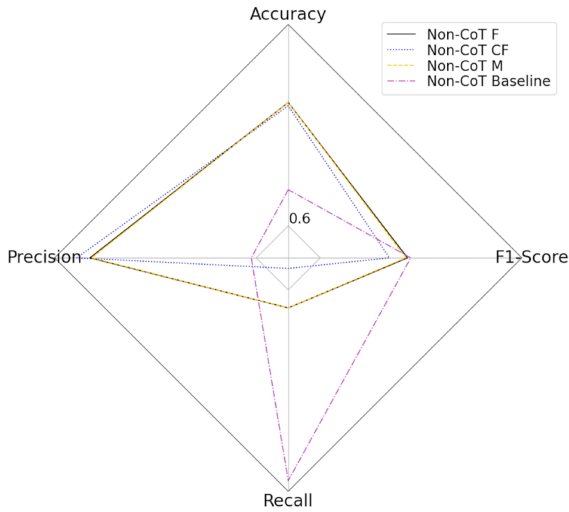
Based on the explanation type, several aspects can be discussed. In the baseline scenarios, explanations are quite brief when discussing the subject, but they become overly verbose when delving into the body of the explanatory narrative. Regarding the remaining scenarios, the purely counterfactual approach tends to yield the shortest narratives (both for the subject and the explanation body), which is surprising as the sets of explanatory factors involved here tend to be larger than those of the purely factual variant. On the other hand, the mixed approach produces the longest explanations, as expected, but their length is far shorter than what would be achieved by the simple concatenation of factual and counterfactual pieces of information, which hints that the LLM is performing an effective summarization of the information received from the FRBSs. This result is promising for the applicability of the proposed method.



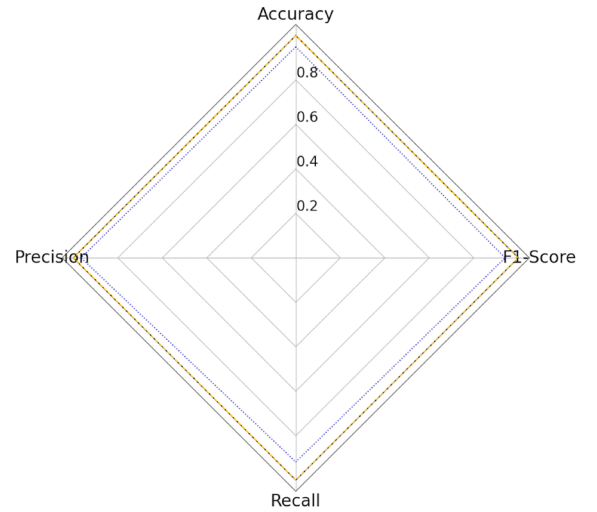
(a) Ground Truth - CoT



(b) Fuzzy Inference - CoT



(c) Ground Truth - Non-CoT



(d) Fuzzy Inference - Non-CoT

Fig. 5. Summary of the predictive performance achieved by the approaches during testing on the diabetes dataset. The first row, i.e., radar plots (a) and (b), shows the values of quality metrics in the scenarios with CoT. The second row, i.e., radar plots (c) and (d), shows the values of the quality metrics for the scenarios without CoT (i.e., Non-CoT). All metrics are reported against the ground truth (plots in the left: a and c) and against the fuzzy inference (plots in the right: b and d). The baseline approach lacks data for the latter.

Regarding the impact of CoT prompting, it increases verbosity across the board, which is not the desired effect. In general, briefer narratives are preferable, in agreement with research findings (see Section 2.3). Thus, unless the increased length is paired with improved readability, CoT may prove counter-productive for the goals of this work. Notice that, the results in the baseline scenarios are mostly unaffected by CoT, which is expected, as it has no initial narrative to structure its discourse around.

Along with the average number of hallucinated attributes (and the standard deviation in brackets) for the scenarios under discussion, attribute skips are also reported in Table 4.

Skipping attributes may occur when dealing with counterfactuals, if a situation arises in which the proposed method is unable to detect hallucinations for a given attribute because the fusion of counterfactuals makes it so any value could be considered correct. As an illustrative example, if the first explainer produced a counterfactual that includes the subject being a heavy drinker, and the second explainer produced a counterfactual deeming the subject a non-heavy drinker, once the LLM fuses the counterfactuals both elements should be on the set of correct factors, which makes our detection method unable to disclose their hallucinatory character based on presence alone.

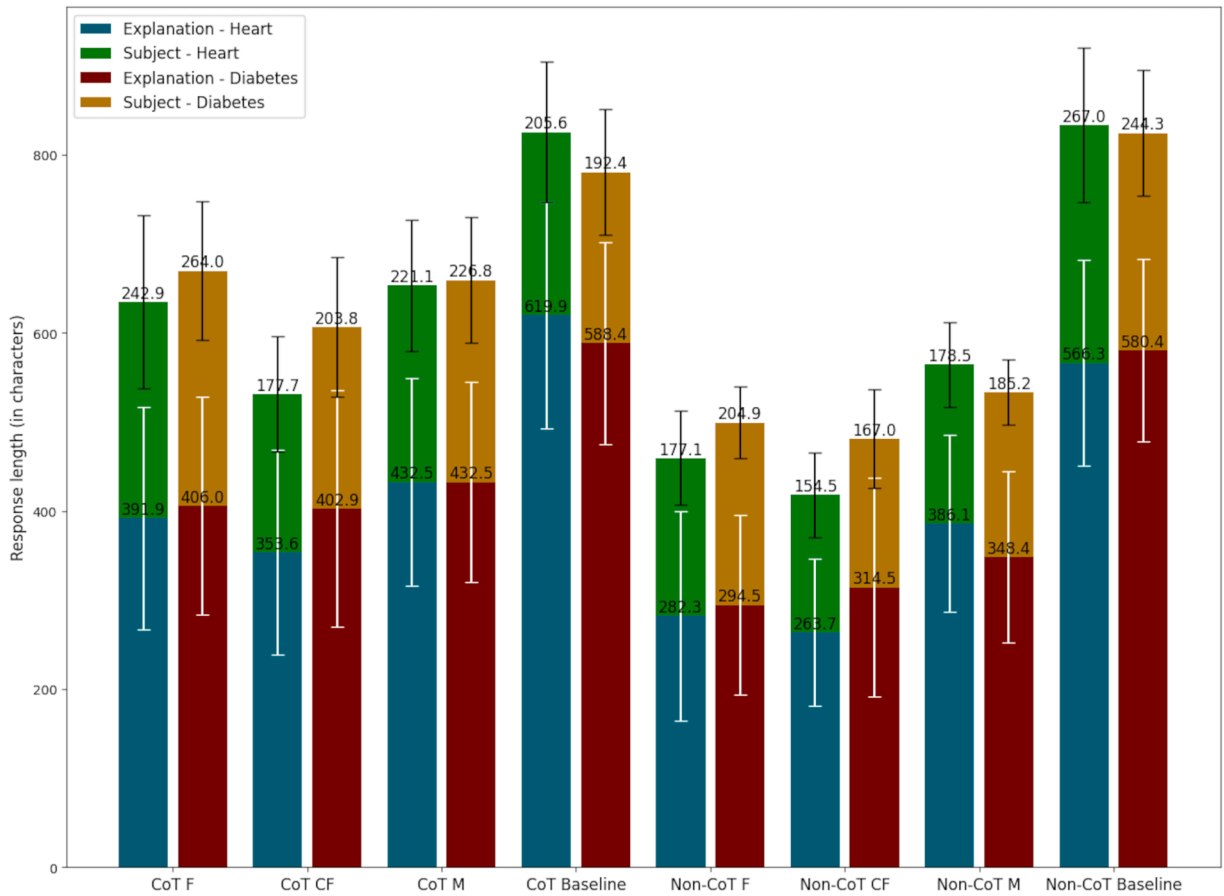


Fig. 6. Summary of response lengths presented by the approaches during testing. For each approach, left bar shows values for the heart disease dataset, right bar shows values for the diabetes dataset. Lower half of bars show explanation lengths, upper half of bars show subject description lengths. Metrics are reported as average response length in characters, with error bars displaying standard deviation.

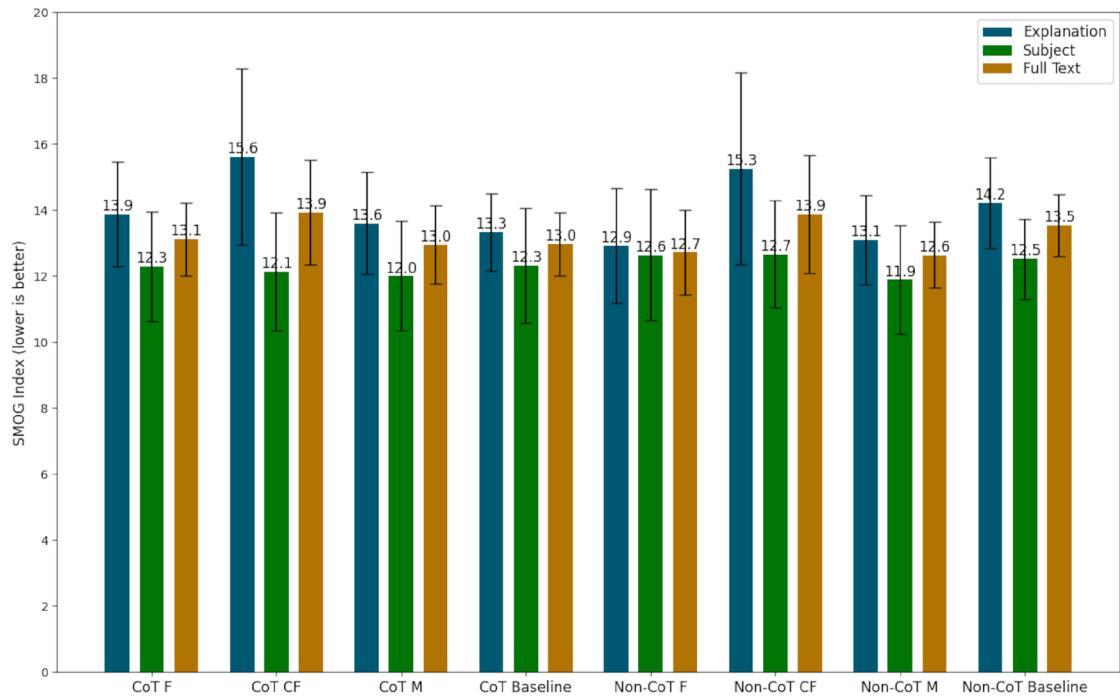
Table 4

Summary of hallucination-related events during testing. “Halluc.” reports detections of hallucinated attributes, “Skips” reports attributes being skipped during detection due to any value potentially being correct. Both hallucinations and skips are reported in averages (with standard deviation in brackets). Factual approaches cannot skip attributes, thus N/A is displayed.

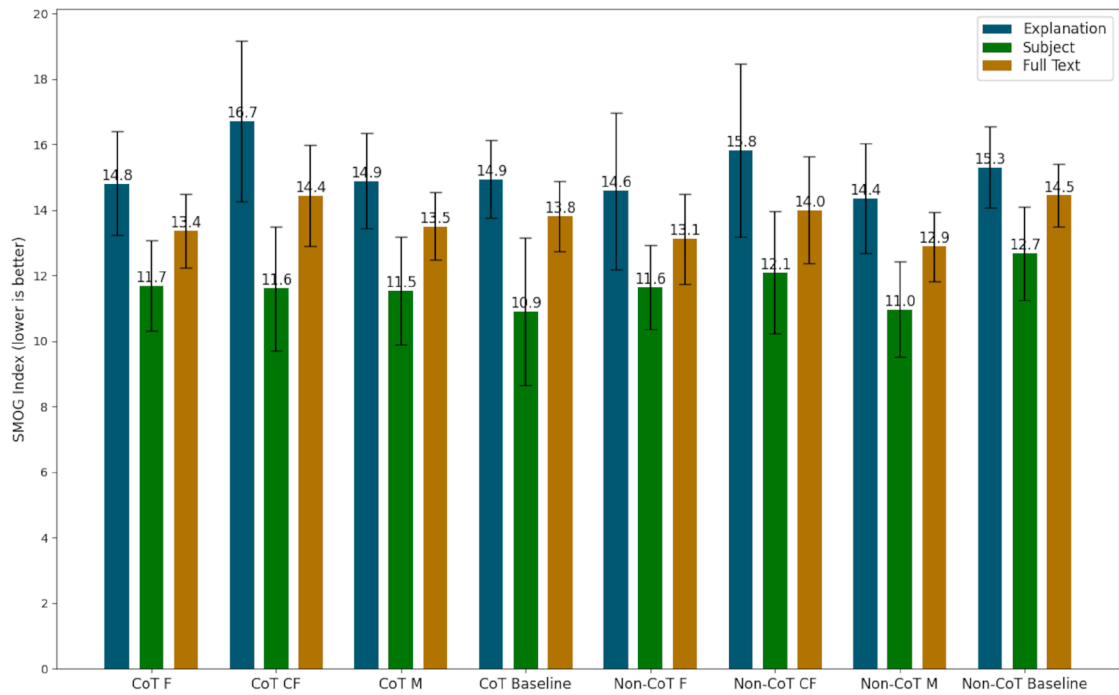
Hallucinations	Heart - Halluc.	Heart - Skips	Diabetes - Halluc.	Diabetes - Skips
Factual + CoT	0.945 (0.902)	N/A	1.268 (0.957)	N/A
Counterfactual + CoT	0.743 (0.873)	0.93 (1.222)	0.94 (0.933)	0.449 (0.6)
Factual	0.631 (0.775)	N/A	0.798 (0.839)	N/A
Counterfactual	0.822 (0.881)	0.93 (1.222)	1.042 (0.916)	0.449 (0.6)

From the reported results, we observe that counterfactuals, even when offered alone, do not induce stronger hallucinatory tendencies than their factual counterpart. They are even more reliable if applying CoT. This may signal that, for a less intuitive type of explanation, the additional reasoning step is more helpful for the LLM. Anyway, the hallucination frequency is, all-around, low: the maximum number attainable for this metric would be 7 (indicating that every possible attribute has been detected to contain potentially hallucinatory information), while most results congregate around the 0.8-1.0 range. In addition, as a brief commentary, attribute skips appear to be infrequent (as seen in Table 4). If they were more common, they could undermine the reliability of hallucination measurements in the counterfactual case, as too many unchecked attributes would skew the results in their favor.

Additionally, Fig. 7 includes two plots: the upper plot corresponds to the heart dataset; the lower plot corresponds to the diabetes dataset. Each scenario is reported with three bars. From left to right, each bar concerns the readability score for: the explanatory narrative (left blue bar), the subject description (central green bar), and the full narrative (i.e., the right orange column). It is worth noting that the full narrative merges the other two pieces of information (i.e., the explanatory narrative and the subject description,



(a) Heart Disease



(b) Diabetes

Fig. 7. Summary of readability scores presented by the different approaches during testing with the two datasets: (a) Heart disease and (b) Diabetes. For each approach, three bars are included, which from left-to-right report readability for the explanation (left blue bars), for the subject description (central green bars), and for the full narrative (right orange bars). Bars are annotated with the average SMOG Index, and error bars display standard deviations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5

Summary of the errors occurring during the test runs. “Errors” report instances of the LLM not complying with its instructed format, “Incons.” report instances of the LLM’s prediction disagreeing with its fuzzy-grounded context. Baseline approaches do not receive said context, thus N/A is displayed.

Errors	Heart - Errors	Heart - Incons.	Diabetes - Errors	Diabetes - Incons.
Factual + CoT	15	0	0	0
Counterfactual + CoT	40	991	18	550
Mixed + CoT	1	2	0	3
Baseline + CoT	64	N/A	100	N/A
Factual	0	0	0	0
Counterfactual	1	665	1	357
Mixed	0	0	0	2
Baseline	1	N/A	1	N/A

but unlike for the average response lengths, the relation between the readability of the full narrative and the scores of its components is not additive, which is why it is explicitly shown.

Given the differences in explanation strategies, the readability of the subject descriptions appears quite uniform across all scenarios, with the major differences observed in the CoT vs. Non-CoT comparison. Additionally, we observe a bigger difference with respect to the readability of the body of explanation: the purely counterfactual approach produces consistently less readable narratives, while the purely factual and mixed approaches produce nearly equal narratives in terms of readability. Moreover, the baseline approach falls in the middle ground, being worse than the factual and mixed approaches, but better than the counterfactual one. Other than the relative performances of baseline versus counterfactual, which were harder to predict, the behavior shown matches expectations. The contrastive nature of counterfactuals makes them ill-suited for independent utterance, resulting in narratives that are more difficult to comprehend.

As for CoT, dual behavior can be observed. It seems beneficial for the baseline approach, as expected, because better text planning and structuring should yield a higher-quality, more readable final utterance. Nevertheless, we obtain worse results with CoT in all the remaining scenarios under study. The effect is slight, but noticeable, and similar to what happened with the length of explanations, contrary to the goals of this work. Accordingly, adding CoT prompting is not helpful for the proposed explanatory strategies, as test results do not improve, and adding that reasoning step only slows the system and makes it less efficient computationally.

As a brief note: the readability results are on the higher end. Even the optimal approach for the full narrative (i.e., a mixed approach without CoT) in both datasets can be rounded to a SMOG Index of 13, indicating that a person would need 13 years of education in order to fully comprehend the text. This roughly means a 1st year college student, which is a significant demand. The baseline being even worse, showcases that LLMs in general seem to have trouble producing plain language in the healthcare domain, which may limit their usefulness for the wider population.

For completeness purposes, Table 5 shows the errors that took place during the experimental runs. Errors refer to instances in which the LLM was unable to generate a response complying to the output format specified in its prompt. Inconsistencies refer to instances in which the LLM’s prediction differed with that of the fuzzy explainers.

When it comes to complying with the required format, we observe bigger problems in the baseline scenarios. The opposite result might seem more intuitive, as this approach is given greater freedom, yet the proposed explanatory strategies streamline the task, allowing the LLM to better comply with additional instructions. It is worth noting that the Non-CoT alternatives produced fewer errors, but their format specification is also simpler, as it does not include a field for storing the reasoning, which skews the comparison. Finally, inconsistencies are extremely low for the mixed approach, with the factual scenario presenting 0 across all its variations. The purely counterfactual scenario shows larger numbers. This aligns with the insights derived from the previous analysis on the reported accuracy.

4.2. Qualitative analysis of generated narratives

To supplement the results of the automatic validation presented in the previous section, we will delve here into a comparative example of the explanatory narratives generated during testing, so that differences between the approaches are apparent. While an exhaustive analysis of all narratives is not feasible given the volume of data, we consider that an illustrative, qualitative approach is informative. In the remainder of this section, we will therefore showcase an example for each test dataset.

4.2.1. Example of narratives for heart disease

Instance 5 of the heart disease dataset was presented by the fuzzy explainer as follows:

This subject has the following data: they are a male, they are of adult age, its BMI is 27.71, they are not a smoker, they are physically active, they are not a heavy drinker and they do not have diabetes.

Keep in mind that this will always be a faithful representation of the subject, as no inference has been performed yet, it only comprises a verbalization of the data point’s attributes. For this instance, all approaches except the baseline with no CoT report the correct predicted class, which is NoHeartDisease.

The baseline with CoT approach, being prompted to reword this subject profile, and explain its potential health status, generated the following narrative:

The subject is a 30-year-old male who is slightly above average weight, with a BMI of 27.71. He does not smoke and is physically active, but his diet and lifestyle may be contributing to his current weight status. He also does not consume excessive amounts of alcohol and does not have diabetes. Based on the subject's profile, there are both risk and protective factors for heart disease. His overweight status increases his risk, but being physically active is beneficial. Not smoking or drinking excessively reduces his risk further. However, without more information about his family history, blood pressure, cholesterol levels, or other health indicators, it is difficult to make a definitive assessment. Nevertheless, considering his age, sex, and BMI, he may be at an increased risk for developing heart disease compared to someone with a normal weight and no other risk factors.

We can observe some concerning factors in the baseline explanations that are not constrained by the fuzzy grounding. Namely, the LLM has already fabricated the subject's age, which was merely stated to be an adult, and added a commentary on diet and lifestyle that is not justified by any of the data at its disposal. In its prediction, it does not make an explicit judgment, like it was asked to. While this caution is certainly valuable, it then proceeds to state that the subject may be at an increased risk for heart disease, which runs contrary to it reporting a negative prediction.

As its more direct comparison, the factual + CoT approach receives the same subject profile from the fuzzy explainer, and additionally, the following fuzzy-grounded explanation:

This person does not present heart disease because they are of adult age or younger and they do not have diabetes. Additional evidence for this is that they are not a smoker.

Following these inputs, the narrative generated by this approach is as follows:

The subject is a male adult with a BMI of 27.71, indicating he is overweight. He does not smoke, is physically active, and consumes alcohol moderately. Importantly, he does not have diabetes. His age places him within the adult range or younger, which is a significant factor in this context. Given his age and lack of diabetes, the subject's risk factors for heart disease are reduced. Furthermore, his non-smoking habit significantly contributes to his lower risk profile. These factors collectively support the conclusion that he does not present with heart disease.

Some of the results shown previously are reflected in this example, most importantly the fuzzy-grounding is helpful for brevity. There is a slight problem with this utterance: the subject is mischaracterized as being of adult age or younger, which is not true as the subject is explicitly stated to be of adult age. This is probably a consequence of the fuzzy explanation, which mentioned adult or younger age as a factor, and this polluted the LLM's reasoning. Otherwise, the explanation works fine: it is concise, does not deviate from the facts presented by the fuzzy explainer, and rewrites the initial narrative into a more natural human-like form.

It may be illustrative to showcase the behavior of the factual approach when no CoT is utilized. After receiving the same inputs as its CoT-enhanced counterpart, the full narrative is as follows:

The subject is a male adult with a BMI of 27.71, who does not smoke, is physically active, and has a moderate drinking habit. He also does not have diabetes. This person was diagnosed as not having heart disease due to their age and lack of diabetes. Additionally, their non-smoking status provided further evidence for this diagnosis.

As shown by the quantitative analysis in [Section 4.1](#), the non-CoT approach produces even more concise responses while remaining remarkably faithful to the information provided by the fuzzy inference. A small consideration to make, which also affects the CoT-enhanced narrative, is that the LLM identifying "not a heavy drinker" with a person who drinks, but moderately, is not a mistake in itself. The datasets do not distinguish between a teetotal subject and one who drinks responsibly, and thus, it is not fair to expect the LLM to make an accurate judgment on this matter.

As for the counterfactual approach, we will comment on the Non-CoT alternative for this example because, according to the quantitative analysis, it usually yields the briefer, more accurate narratives. The subject description is the same we have been working with, but the related fuzzy explanation is different:

This person would present heart disease IF they were a middle-aged person, they were not a smoker and they were a male AND EITHER: 1) they were physically active, their BMI did not indicate obesity of type 3 and they had diabetes ; OR 2) their BMI did not indicate underweight, they were not a heavy drinker and they did not have diabetes.

The resulting narrative is as follows:

This male adult has a BMI of 27.71, indicating he is overweight but not obese. He does not smoke, is physically active, and consumes alcohol in moderation. Importantly, he does not have diabetes. For this individual to develop heart disease, two scenarios are possible: Scenario 1 - If he were middle-aged, non-smoker, and male, his risk would increase if he remained physically active, had a normal weight (not underweight or obese), and also had diabetes. Scenario 2 - Alternatively, if he were middle-aged, non-smoker, and male, his risk would increase if he was not physically active, did not have obesity type 3, and did not have diabetes.

The subject description is essentially the same as in the factual case, as expected. However, the explanation is of lower quality overall: while the first scenario is mostly faithful to the fuzzy inference (there is only a minor mishap in mentioning normal weight, when the grounded explanation merely says that it should not be type 3 obesity), the second scenario gets the weight-related factor wrong, probably a mix up from scenario 1, and substitutes the drinking-related factor for an uncalled-for mention of physical activity. These problems point to the increased complexity of counterfactuals, which confound the LLM during the rewriting process.

Finally, the mixed approach (again, Non-CoT) takes the following template-based explanation provided by the fuzzy explainer (as it may be evident, this is a concatenation of the factual and counterfactual explanations, although avoiding repetitions between them):

This person does not present heart disease because they are of adult age or younger and they do not have diabetes. Additional evidence for this is that they are not a smoker. This person would present heart disease IF they were a middle-aged person and they were a male AND EITHER: 1) they were physically active, their BMI did not indicate obesity of type 3 and they had diabetes; OR 2) their BMI did not indicate underweight and they were not a heavy drinker.

Then, the LLM-rewritten narrative is:

The subject is a male adult with a BMI of 27.71, indicating he is overweight but not obese. He does not smoke, is physically active, and consumes alcohol in moderation. Additionally, he has been tested for diabetes and does not have it. This person's age and lack of diabetes are strong indicators that they do not have heart disease. Furthermore, their non-smoking habit also reduces the risk of heart disease. However, if this individual were middle-aged and male, and either had diabetes or was physically active with a normal weight and did not consume excessive amounts of alcohol, they would be at higher risk for developing heart disease.

This approach, logically, seems to rest on a middle ground between the factual and counterfactual results. Up to the end of the factual explanation, there are few to no problems, but some issues arise when the counterfactual information begins: the breakdown of the scenarios is incorrect, mixing information as the purely counterfactual approach also did. As a slight advantage over the latter, though, it did not completely fabricate anything.

4.2.2. Example of narratives for diabetes

Instance 361 of the diabetes dataset was presented by the fuzzy explainer as follows:

This subject has the following data: they are a male, they are of adult age, its BMI is 29.0, they are a smoker, they are not physically active, they are not a heavy drinker and they do not have heart disease.

For this instance, all the fuzzy-grounded approaches correctly predict NoDiabetes, whereas the baselines do not. For completeness, and given that we previously studied mostly the non-CoT variants of the explanations, we will now pay more attention to their CoT equivalents, despite the validation results showing they are inferior. Starting with the CoT baseline, the LLM-based rewriting of the subject plus prediction explanation is:

The subject is a 29-year-old male adult with a BMI of 29.0. He is a smoker and does not engage in regular physical activity. He also does not have heart disease. His lifestyle habits suggest that he may be at risk for various health issues, including diabetes. Based on the subject's profile, his high BMI and lack of physical activity increase his risk of developing insulin resistance, a precursor to type 2 diabetes. Additionally, smoking can further exacerbate this risk by damaging blood vessels and nerves, making it more difficult for glucose to enter cells. While there are no other medical conditions mentioned, such as heart disease, which could be related to diabetes, the subject's current lifestyle habits pose a significant threat to his metabolic health. Therefore, it is essential for him to adopt a healthier lifestyle, including regular exercise and a balanced diet, to reduce his risk of developing diabetes.

This utterance is quite similar to the heart disease equivalent for this approach: it is very long, fabricates information (e.g., the subject's age), and does not utter a prediction (as it was asked to). However, in this case, at least, the arguments put forward by the model are coherent with its reported prediction, which is affirmative.

Continuing to the factual explanation, with CoT prompting, the fuzzy-grounded explanation alone is:

This person does not present diabetes because their BMI indicates overweight, they do not have heart disease and they are of adult age.

Putting together this explanation, along with the subject description, the LLM produces the following explanatory narrative:

This adult male has several key characteristics that are relevant to his overall health. He is overweight, with a BMI of 29.0, which puts him at risk for various health issues. As a smoker, he is exposing himself to harmful chemicals that can lead to serious health problems. Additionally, his sedentary lifestyle means he does not engage in regular physical activity, which is essential for maintaining good health. On the other hand, he does not have heart disease and is not a heavy drinker, which are both positive factors in his health profile. Based on the subject's characteristics, particularly his BMI indicating overweight status, lack of heart disease, and adult age, it was determined that he does not present diabetes. This conclusion is drawn from the medical assessment of his overall health profile.

We observe that the explanation part of this utterance is good: short and to the point, without deviating from the information that was provided to the model. However, the subject description is too long and overly complicated: the LLM adds a lot of content that was not present in the original narrative. Although the new content is accurate in this case, it still undermines the trustworthiness of the system because uncontrolled information is leaking into the final narratives.

Despite focusing on CoT narratives for this illustrative example, it is probably best to check the Non-CoT counterpart to the previous explanation to see if the over-explanation problem on the subject is fixed. In this case, the LLM takes the same inputs and produces the following output:

The subject is a male adult with a BMI of 29.0, indicating he is overweight. He has a history of smoking but does not engage in regular physical activity. Additionally, he does not consume excessive amounts of alcohol and does not have heart disease. This individual was diagnosed as not having diabetes due to his BMI being within the overweight range, the absence of heart disease, and his age being considered adult. These factors collectively led to this conclusion.

As expected, in light of the experimental results already discussed in [Section 4.1](#), the Non-CoT alternative does create a briefer, less convoluted equivalent of the explanation without loss of accuracy.

Returning to the CoT-enhanced approaches, the purely counterfactual variant takes the same subject description, but the fuzzy explainer produces a different template-based explanation:

This person would present diabetes IF they were a smoker, their BMI indicated obesity of type 1, they were not physically active, they did not have heart disease and they were of adult age.

Then, the LLM rewrites the utterance as follows:

The subject is a 29-year-old male with a BMI of 29.0, indicating he is overweight. He has a sedentary lifestyle, which means he does not engage in regular physical activity. Additionally, he smokes, but he does not have heart disease or drink heavily. For this person to develop diabetes, several factors would need to change. Firstly, his smoking habit would remain unchanged. Secondly, his BMI would need to indicate obesity of type 1, meaning he would need to gain more weight. Thirdly, his lack of physical activity would continue to be a contributing factor. Lastly, the absence of heart disease and being an adult would still apply.

We can see that, because this counterfactual is simpler than the one presented for the heart disease instance, this approach now works better. The explanation is accurate to the explanatory facts supplied by the fuzzy interface, reasonably brief, and contains only a minor mistake: the LLM has fabricated the subject's age (again, as it also happened for the baseline). Interestingly, the LLM even adds some stylistic choices to point out the many of the factors already held for the subject (e.g., "would still apply", "would remain unchanged").

Finally, the CoT-enhanced mixed approach takes the usual combination of factual and counterfactual fuzzy inferences, without repetition:

This person does not present diabetes because their BMI indicates overweight, they do not have heart disease and they are of adult age. This person would present diabetes IF they were not physically active, their BMI indicated obesity of type 1 and they were a smoker.

And, in turn, outputs the following mixed narrative:

The subject is a male adult with a BMI of 29.0, indicating he is overweight. He has a sedentary lifestyle, being neither physically active nor a heavy drinker. Additionally, he smokes, but does not have heart disease. Based on his profile, this person would be at risk for developing diabetes if he were more inactive, had a higher BMI indicating obesity type 1, and was a smoker. However, given his current situation, he does not present with diabetes due to his age, lack of heart disease, and relatively low BMI compared to the threshold for obesity type 1.

Similarly to what happened for the purely counterfactual approach, the simplicity of this instance compared to the one in the heart disease dataset allowed the mixed scenario to improve its performance: the subject description is short and correct (although there is a mishap, relating absence of heavy drinking with a sedentary lifestyle), and both parts of the explanation are faithful to the fuzzy-grounded information too. Interestingly, the LLM decided to reorder the explanation, putting counterfactuals first, factuals second.

5. Conclusions and future work

The three proposed explanatory strategies (i.e., factual and/or counterfactual approaches) are overall beneficial for trustworthiness. Fuzzy explainers display predictive performance that aligns well with the underlying tree models, especially for the factual and mixed cases, while producing brief explanations with a good level of readability (compared to the purely LLM-based baseline). Moreover, the small number of reported errors shows that LLMs are quite adept at following the instructions needed to couple them with the template-based fuzzy-grounded explanatory interface that takes as input the factual and counterfactual information provided by FRBSs.

Accordingly, we validated the first research hypothesis (H1) in light of the reported results. While the purely counterfactual approach was not a strong performer compared to its counterparts, the mixed mode of explanations (i.e., the combination of factual and counterfactual explanations) stood out as the best explanatory strategy. It showcased equal predictive performance to the purely factual strategy, with a comparatively small increase in overall length that does not result in a less readable text. Given that neither the factual nor the counterfactual approaches showed a strong tendency towards hallucination, it is reasonable to conclude that the mixed mode of explanations is not making the explainer less reliable either, while offering richer narratives that fuse the actual prediction with its contrastive hypotheticals.

On the contrary, the second hypothesis (H2) could not be validated empirically. The explanatory strategies enriched with CoT prompting did not exhibit the expected relevant advantages over their non-CoT peers. Indeed, they produced longer, less readable narratives, thus being counterproductive to the goals of this work. This behavior is only reverted in the baseline scenario.

It is worth noting that the failure of CoT prompting is not particularly rare. For example, CoT is documented to fail in planning domains [73]: testing over the classical Blocksworld problem, along with an additional set of non-planning problems scalable in complexity, showed that CoT does not always induce strong generalization capabilities, and that its effectiveness is highly dependent on prompt specificity, being reliant on sufficiently detailed examples that enable the model to solve its task by pattern matching.

We applied "zero-shot" CoT prompting, i.e., we only instruct the model to reason about its task without providing specific examples, and this lack of specificity could explain why the CoT prompting was not yielding good results in our experiments.

In addition, Shojaei et al. [74] researched Large Reasoning Models, i.e., LLMs applied to reasoning tasks. They found that reasoning models were not always better than non-reasoning equivalents for simple tasks, but performed better on medium-complexity tasks before collapsing again in the hardest problems. This could hint at why, in our empirical quantitative analysis, CoT prompting was only slightly effective in improving the baseline approach, which lacks additional information from the fuzzy explainer and thus faces a medium-complexity task. However, CoT prompting was not effective in the easier task, i.e., when the information was preprocessed by the LLM in combination with the fuzzy explainer.

All in all, we consider counterfactuals a beneficial addition to the general strategy of merging FRBSs with LLMs, especially when using the mixed mode, where counterfactuals are provided alongside factual pieces of information. However, the utility of CoT prompting for the rewriting task is limited under the experimental conditions considered.

This work has a few limitations worth noting. Firstly, the information fusion strategy considers only the two winner rules in the two given FRBSs. Secondly, the validation was performed on two closely linked datasets, which limits the generalizability of the findings to more diverse domains. Thirdly, the hallucination detection methodology that was utilized must be further improved to be more accurate in its operation and more broadly applicable. By the way, it is worth mentioning that this is an open problem, especially when it is undesirable to involve other LLMs in the process for the sake of simplicity and trustworthiness. Finally, the evaluation framework relies solely on automatic metrics: while this serves as a first-pass assessment of the validity of the proposal, the quantitative analysis of trustworthiness and explanatory quality should be complemented in the future with a qualitative analysis grounded in a series of user studies.

Accordingly, our future research agenda is motivated by the identified limitations. In the long term, we plan to extend the information fusion strategy to more than two FRBSs, and to more than a single winner rule in each FRBS (i.e., considering the k-top fired rules instead of only the winner rule when elaborating factual and counterfactual explanations). In addition, we will generalize the validation of the explanatory strategy and interface across a wider range of domains. We will also refine the hallucination detection methodology. Moreover, we will carefully design a series of user studies to further investigate the CoT's effectiveness and the correlation between automatic metrics and human evaluation.

Finally, for the sake of open science and the reproducibility of experiments, all resources are released in the following GitLab project: <https://gitlab.citius.gal/pablomiguel.perez/fcflm-citius-ijar>

CRedit authorship contribution statement

Pablo Miguel Perez-Ferreiro: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Jose M. Alonso-Moral:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization; **Alejandro Catala:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization; **Alberto Bugarín-Diz:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Funding

Pablo Miguel Perez-Ferreiro is a PhD Researcher supported by the Galician Ministry of Education, Science, Universities and Professional Training (ED481A-2026-056). All authors recognize the support of the Galician Ministry of Education, Science, Universities and Professional Training (grants CITIUS 2024-2027 ED431G2023/04 and GSI 2026-2029 ED431C 2026/52) and the European Union (European Regional Development Fund - ERDF). This work is also supported by the Spanish Ministry of Science, Innovation and Universities (MCIN/AEI/10.13039/501100011033/) through MAIXAI4TRUST PID2024-157680NB-I00 and COLLAGE PID2023-149549NB-I00 grants and the “Fundación Ramón Areces” (Spain) through the CONFIA project.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jose M. Alonso-Moral reports financial support was provided by Ramon Areces Foundation. Given his role as Area Editor, Alberto Bugarín-Diz had no involvement in the peer review of this article and had no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to another journal editor. Given his role as Associate Editor, Jose M. Alonso-Moral had no involvement in the peer review of this article and had no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to another journal editor. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the link to the data/code in the revised manuscript.

Appendix A. Rule fusion example

In this appendix, an end-to-end example is given for the rule fusion procedure that was presented in Section 3.2. Keep in mind that the notation introduced in that section is used here for clarity.

Let's consider the very first instance of the heart disease dataset as an illustrative example. In this instance, both fuzzy explainers (EXP^1 and EXP^2) agree on predicting the presence of heart disease. The premises in the winner rules are as follows:

$$\begin{aligned} p^1 &= \text{Gender is Female AND Diabetic is Yes AND Age is MiddleAged AND Smoker is Yes AND PhysicalActivity is No} \\ p^2 &= \text{Diabetic is Yes AND Smoker is Yes AND Age is MiddleAged} \end{aligned}$$

Computing the relevant sets for the explanation fusion would result in:

$$\begin{aligned} UF^1 &= \{\text{Gender is Female, PhysicalActivity is No}\} \\ UF^2 &= \{\emptyset\} \\ SF &= \{\text{Diabetic is Yes, Smoker is Yes, Age is MiddleAged}\} \\ AF &= \{\emptyset\} \end{aligned}$$

Following the rules given on Section 3.2.2, because SF is non-empty, it will be verbalized as the main contributor to the explanation, with the UF sets being presented as supporting evidence. According to this, the final explanation that the fuzzy explainer would produce is:

“This person presents heart disease because they are a middle-aged person, they are a smoker and they have diabetes. Additional evidence for this is that they are not physically active and they are a female.”

Now for the counterfactual case on this same instance, both EXP^1 and EXP^2 have a potential explanation to offer. Their premises are:

$p^{c1} = \text{BodyMassIndex is (UnderWeight OR normalWeight) AND Gender is Female AND Diabetic is Yes AND Smoker is Yes AND Age is MiddleAged AND PhysicalActivity is No}$

$p^{c2} = \text{BodyMassIndex is NotObesity AND Gender is Female AND Diabetic is No AND Smoker is Yes AND Age is MiddleAged AND PhysicalActivity is Yes}$

We can now compute the relevant sets:

$UF^{c1} = \{\text{BodyMassIndex is UnderWeight OR normalWeight, Diabetic is Yes, PhysicalActivity is No}\}$

$UF^{c2} = \{\text{BodyMassIndex is NotObesity, Diabetic is No, PhysicalActivity is Yes}\}$

$SF^c = \{\text{Gender is Female, Smoker is Yes, Age is MiddleAged}\}$

$AF^c = \{(\text{BodyMassIndex is UnderWeight OR normalWeight, BodyMassIndex is NotObesity}), (\text{Diabetic is No, Diabetic is Yes}), (\text{PhysicalActivity is No, PhysicalActivity is Yes})\}$

And following the rules for crafting the counterfactual explanation, we could produce two different explanatory narratives. For the stand-alone mode, in which repetition of information already included on the factual case is acceptable, we would obtain:

“This person would not present heart disease IF they were a middle-aged person, they were a smoker and they were a female AND EITHER: 1) they were not physically active, their BMI indicated underweight or normal weight and they had diabetes ; OR 2) they were physically active, their BMI did not indicate obesity and they did not have diabetes.”

For the supplementary mode, avoiding repetition, we would end up with:

“This person would not present heart disease IF their BMI indicated underweight or normal weight OR IF they were physically active, their BMI did not indicate obesity and they did not have diabetes.”

Appendix B. Sample prompts

In this appendix, all the prompt templates utilized during the experimental phase of this work are reported. Note that elements in curly brackets are substituted by the appropriate information when the prompt is sent to the LLM.

Chain-Of-Thought / Mixed and Factual

```

You are an expert AI assistant intended to help your user
  learn about {illness}.

You will be provided with the following data:
--A subject's profile: including his data.
--The subject's diagnosis: which includes the presence or
  lack thereof for {illness} along an explanation of the
  diagnosis.

The data is the following:
--Subject profile: {subject}
--Subject's diagnosis: {diagnosis}

Your task is to REWRITE the subject's profile and diagnosis
  in such a way that they form a COHERENT EXPLANATORY PIECE.
You MUST NOT alter the elements of the subject or the
  diagnosis in a meaningful way: you may only rewrite them
  into a more HUMANLIKE form.
You MUST NOT add additional DETAILS not included in your
  supplied data.
You MUST ASCERTAIN the CURRENT HEALTH STATUS of the subject
  regarding the {illness}.
The resulting EXPLANATIONS should be FAITHFUL to the data
  supplied to you, and HELPFUL to learn about the illness
  and the subject.

DO NOT WRITE THE EXPLANATIONS DIRECTLY: First, reason what
  you NEED TO WRITE, complying with your instructions, and
  then WRITE THE EXPLANATIONS.

Provide your final answer as a JSON with FOUR FIELDS, ALL OF
  THEM IN PLAIN TEXT:
The field 'reasoning' should contain a PLAIN TEXT of your
  FULL REASONING as to WHAT YOU NEED TO WRITE.
The field 'subject' should contain a PLAIN TEXT of your
  REWRITING OF THE SUBJECT PROFILE.
The field 'explanation' should contain a PLAIN TEXT of your
  REWRITING OF THE DIAGNOSIS.
The field 'diagnosis' should contain EXCLUSIVELY either '1.0'
  or '2.0', depending on the CURRENT HEALTH STATUS, where
  1.0 indicates that THERE IS NO ILLNESS, and 2.0 indicates
  that THERE IS ILLNESS.

```

Chain-Of-Thought / Counterfactual

You are an expert AI assistant intended to help your user learn about {illness}.

You will be provided with the following data:

- A subject's profile: including his data.
- The subject's counterfactual diagnosis: which describes what factors of the subject's profile would need to change so that his CURRENT HEALTH STATUS for {illness} CHANGED to a NEW HEALTH STATUS (for better or worse).

The data is the following:

- Subject profile: {subject}
- Subject's counterfactual diagnosis: {diagnosis}

Your task is to REWRITE the subject's profile and counterfactual diagnosis in such a way that they form a COHERENT EXPLANATORY PIECE.

You MUST NOT alter the elements of the subject or the counterfactual diagnosis in a meaningful way: you may only rewrite them into a more HUMANLIKE form.

You MUST NOT add additional DETAILS not included in your supplied data.

You MUST ASCERTAIN the current health status, but you MUST NOT DISCUSS IT or ITS CAUSES: TALK EXCLUSIVELY about the elements included in the COUNTERFACTUAL DIAGNOSIS.

The resulting EXPLANATIONS should be FAITHFUL to the data supplied to you, and HELPFUL to learn about the subject and how the subject's health may CHANGE.

DO NOT WRITE THE EXPLANATIONS DIRECTLY: First, reason what you NEED TO WRITE, complying with your instructions, and then WRITE THE EXPLANATIONS.

Provide your final answer as a JSON with FOUR FIELDS, ALL OF THEM IN PLAIN TEXT:

- The field 'reasoning' should contain a PLAIN TEXT of your FULL REASONING as to WHAT YOU NEED TO WRITE.
- The field 'subject' should contain a PLAIN TEXT of your REWRITING OF THE SUBJECT PROFILE.
- The field 'explanation' should contain a PLAIN TEXT of your REWRITING OF THE COUNTERFACTUAL DIAGNOSIS.
- The field 'diagnosis' should contain EXCLUSIVELY either '1.0' or '2.0', depending on the CURRENT HEALTH STATUS (not the NEW HEALTH STATUS), where 1.0 indicates that THERE IS NO ILLNESS, and 2.0 indicates that THERE IS ILLNESS.

Chain-Of-Thought / Baseline

You are an expert AI assistant intended to help your user learn about {illness}.

You will be provided with the following data:
 --A subject's profile: including his data.

The data is the following:
 --Subject profile: {subject}

Your task is to ASCERTAIN the subject's CURRENT HEALTH STATUS regarding {illness} and CREATE A COHERENT EXPLANATORY PIECE that EXPLAINS your DIAGNOSIS BASED ON THE SUBJECT'S PROFILE.

You MUST NOT alter the elements of the subject's profile in a meaningful way: you may only rewrite them into a more HUMANLIKE form.

The resulting EXPLANATIONS should be FAITHFUL to the PROFILE supplied to you, and HELPFUL to learn about the illness and its risk factors.

DO NOT WRITE THE EXPLANATIONS DIRECTLY: First, reason what you NEED TO WRITE, complying with your instructions, and then WRITE THE EXPLANATIONS.

Provide your final answer as a JSON with FOUR FIELDS, ALL OF THEM IN PLAIN TEXT:

The field 'reasoning' should contain a PLAIN TEXT of your FULL REASONING as to WHAT YOU NEED TO WRITE.

The field 'subject' should contain a PLAIN TEXT of your REWRITING OF THE SUBJECT PROFILE.

The field 'explanation' should contain a PLAIN TEXT of your YOUR EXPLANATION OF THE DIAGNOSIS BASED ON THE PROFILE.

The field 'diagnosis' should contain EXCLUSIVELY either '1.0' or '2.0', depending on the CURRENT HEALTH STATUS, where 1.0 indicates that THERE IS NO ILLNESS, and 2.0 indicates that THERE IS ILLNESS.

No Chain-Of-Thought / Mixed and Factual

You are an expert AI assistant intended to help your user learn about {illness}.

You will be provided with the following data:

- A subject's profile: including his data.
- The subject's diagnosis: which includes the presence or lack thereof for {illness} along an explanation of the diagnosis.

The data is the following:

- Subject profile: {subject}
- Subject's diagnosis: {diagnosis}

Your task is to REWRITE the subject's profile and diagnosis in such a way that they form a COHERENT EXPLANATORY PIECE. You MUST NOT alter the elements of the subject or the diagnosis in a meaningful way: you may only rewrite them into a more HUMANLIKE form.

You MUST NOT add additional DETAILS not included in your supplied data.

You MUST ASCERTAIN the CURRENT HEALTH STATUS of the subject regarding the {illness}.

The resulting EXPLANATIONS should be FAITHFUL to the data supplied to you, and HELPFUL to learn about the illness and the subject.

Provide your final answer as a JSON with THREE FIELDS, ALL OF THEM IN PLAIN TEXT:

The field 'subject' should contain a PLAIN TEXT of your REWRITING OF THE SUBJECT PROFILE.

The field 'explanation' should contain a PLAIN TEXT of your REWRITING OF THE DIAGNOSIS.

The field 'diagnosis' should contain EXCLUSIVELY either '1.0' or '2.0', depending on the CURRENT HEALTH STATUS, where 1.0 indicates that THERE IS NO ILLNESS, and 2.0 indicates that THERE IS ILLNESS.

No Chain-Of-Thought / Counterfactual

You are an expert AI assistant intended to help your user learn about {illness}.

You will be provided with the following data:

- A subject's profile: including his data.
- The subject's counterfactual diagnosis: which describes what factors of the subject's profile would need to change so that his CURRENT HEALTH STATUS for {illness} CHANGED to a NEW HEALTH STATUS (for better or worse).

The data is the following:

- Subject profile: {subject}
- Subject's counterfactual diagnosis: {diagnosis}

Your task is to REWRITE the subject's profile and counterfactual diagnosis in such a way that they form a COHERENT EXPLANATORY PIECE.

You MUST NOT alter the elements of the subject or the counterfactual diagnosis in a meaningful way: you may only rewrite them into a more HUMANLIKE form.

You MUST NOT add additional DETAILS not included in your supplied data.

You MUST ASCERTAIN the current health status, but you MUST NOT DISCUSS IT or ITS CAUSES: TALK EXCLUSIVELY about the elements included in the COUNTERFACTUAL DIAGNOSIS.

The resulting EXPLANATIONS should be FAITHFUL to the data supplied to you, and HELPFUL to learn about the subject and how the subject's health may CHANGE.

Provide your final answer as a JSON with THREE FIELDS, ALL OF THEM IN PLAIN TEXT:

The field 'subject' should contain a PLAIN TEXT of your REWRITING OF THE SUBJECT PROFILE.

The field 'explanation' should contain a PLAIN TEXT of your REWRITING OF THE COUNTERFACTUAL DIAGNOSIS.

The field 'diagnosis' should contain EXCLUSIVELY either '1.0' or '2.0', depending on the CURRENT HEALTH STATUS (not the NEW HEALTH STATUS), where 1.0 indicates that THERE IS NO ILLNESS, and 2.0 indicates that THERE IS ILLNESS.

No Chain-Of-Thought / Baseline

You are an expert AI assistant intended to help your user learn about {illness}.

You will be provided with the following data:
--A subject's profile: including his data.

The data is the following:
--Subject profile: {subject}

Your task is to ASCERTAIN the subject's CURRENT HEALTH STATUS regarding {illness} and CREATE A COHERENT EXPLANATORY PIECE that EXPLAINS your DIAGNOSIS BASED ON THE SUBJECT'S PROFILE.

You MUST NOT alter the elements of the subject's profile in a meaningful way: you may only rewrite them into a more HUMANLIKE form.

The resulting EXPLANATIONS should be FAITHFUL to the PROFILE supplied to you, and HELPFUL to learn about the illness and its risk factors.

Provide your final answer as a JSON with THREE FIELDS, ALL OF THEM IN PLAIN TEXT:

The field 'subject' should contain a PLAIN TEXT of your REWRITING OF THE SUBJECT PROFILE.

The field 'explanation' should contain a PLAIN TEXT of your YOUR EXPLANATION OF THE DIAGNOSIS BASED ON THE PROFILE.

The field 'diagnosis' should contain EXCLUSIVELY either '1.0' or '2.0', depending on the CURRENT HEALTH STATUS, where 1.0 indicates that THERE IS NO ILLNESS, and 2.0 indicates that THERE IS ILLNESS.

Appendix C. Supplementary details on reported results

In this appendix, the full numerical values for all the metrics taken during the validation of this work are reported. [Tables C.6–C.15](#) provide readers with complementary information to the results reported in [Section 4.1](#).

C.1. Performance metrics: accuracy, precision, recall, F1-score

Table C.6

Summary of accuracies achieved during test runs, after LLM processing. “Truth” reports the metric taken against the ground truth of the test set, “FRBS” reports the metric taken against the original fuzzy predictions.

Accuracy	Heart - Truth	Heart - FRBS	Diabetes - Truth	Diabetes - FRBS
Factual + CoT	0.712	1	0.675	1
Counterfactual + CoT	0.651	0.818	0.648	0.922
Mixed + CoT	0.712	1	0.674	1
Baseline + CoT	0.619	N/A	0.612	N/A
Factual	0.712	1	0.675	1
Counterfactual	0.669	0.879	0.673	0.949
Mixed	0.712	1	0.675	1
Baseline	0.617	N/A	0.622	N/A

Table C.7

Summary of precisions achieved during test runs, after LLM processing. “Truth” reports the metric taken against the ground truth of the test set, “FRBS” reports the metric taken against the original fuzzy predictions.

<i>Precision</i>	Heart - Truth	Heart - FRBS	Diabetes - Truth	Diabetes - FRBS
Factual + CoT	0.706	1	0.701	1
Counterfactual + CoT	0.727	1	0.705	0.996
Mixed + CoT	0.706	0.999	0.700	0.999
Baseline + CoT	0.592	N/A	0.641	N/A
Factual	0.706	1	0.701	1
Counterfactual	0.715	0.998	0.709	0.965
Mixed	0.706	1	0.700	0.999
Baseline	0.586	N/A	0.603	N/A

Table C.8

Summary of recall rates achieved during test runs, after LLM processing. “Truth” reports the metric taken against the ground truth of the test set, “FRBS” reports the metric taken against the original fuzzy predictions.

<i>Recall</i>	Heart - Truth	Heart - FRBS	Diabetes - Truth	Diabetes - FRBS
Factual + CoT	0.726	1	0.611	1
Counterfactual + CoT	0.481	0.644	0.508	0.824
Mixed + CoT	0.727	1	0.611	1
Baseline + CoT	0.782	N/A	0.518	N/A
Factual	0.727	1	0.611	1
Counterfactual	0.564	0.765	0.587	0.918
Mixed	0.727	1	0.611	1
Baseline	0.8	N/A	0.716	N/A

Table C.9

Summary of F1-scores achieved during test runs, after LLM processing. “Truth” reports the metric taken against the ground truth of the test set, “FRBS” reports the metric taken against the original fuzzy predictions.

<i>F1-Score</i>	Heart - Truth	Heart - FRBS	Diabetes - Truth	Diabetes - FRBS
Factual + CoT	0.716	1	0.653	1
Counterfactual + CoT	0.579	0.783	0.591	0.902
Mixed + CoT	0.716	0.999	0.652	0.999
Baseline + CoT	0.674	N/A	0.573	N/A
Factual	0.716	1	0.653	1
Counterfactual	0.631	0.866	0.642	0.941
Mixed	0.716	1	0.652	0.999
Baseline	0.676	N/A	0.655	N/A

C.2. Interpretability metrics: Length, hallucinations, readability

Table C.10

Summary of response lengths for the narratives produced by the system, with “Subject” concerning the instance description only and “Expl.” concerning the explanatory piece only. Reported as the average character length the textual piece, with standard deviation in brackets.

<i>Length</i>	Heart - Subject	Heart - Expl.	Diabetes - Subject	Diabetes - Expl.
Factual + CoT	242.949 (97.009)	391.861 (125.068)	263.958 (78.192)	405.993 (122.366)
Counterfactual + CoT	177.735 (65.244)	353.602 (115.071)	203.824 (78.739)	402.897 (132.438)
Mixed + CoT	221.136 (73.727)	432.509 (116.984)	226.788 (70.761)	432.539 (111.992)
Baseline + CoT	205.645 (78.868)	619.918 (127.112)	192.362 (70.794)	588.419 (113.405)
Factual	177.119 (52.712)	282.282 (117.657)	204.926 (39.95)	294.462 (101.195)
Counterfactual	154.539 (47.611)	263.693 (82.529)	166.961 (55.641)	314.473 (122.508)
Mixed	178.459 (47.389)	386.086 (98.874)	185.221 (36.697)	348.419 (96.667)
Baseline	266.966 (86.812)	566.305 (115.359)	244.257 (70.225)	580.364 (102.642)

Table C.11

Summary of response lengths for the full-text narratives produced by the system, combining description and explanation, reported as the average character length of the entire piece, with standard deviation in brackets.

<i>Length</i>	Heart - Full Text	Diabetes - Full Text
Factual + CoT	634.81 (185.173)	669.951 (169.113)
Counterfactual + CoT	531.336 (155.012)	606.721 (186.522)
Mixed + CoT	653.645 (162.305)	659.327 (150.084)
Baseline + CoT	825.563 (139.188)	780.781 (137.378)
Factual	459.4 (150.937)	499.388 (119.973)
Counterfactual	418.232 (114.812)	481.435 (161.939)
Mixed	564.545 (121.282)	533.64 (106.954)
Baseline	833.271 (123.693)	824.622 (118.504)

Table C.12

Summary of hallucination-related events during testing. “Halluc.” reports detections of hallucinated attributes, “Skips” reports attributes being skipped during detection due to any value potentially being correct. Reported as averages, plus standard deviation in brackets.

<i>Hallucinations</i>	Heart - Halluc.	Heart - Skips	Diabetes - Halluc.	Diabetes - Skips
Factual + CoT	0.945 (0.902)	N/A	1.268 (0.957)	N/A
Counterfactual + CoT	0.743 (0.873)	0.93 (1.222)	0.94 (0.933)	0.449 (0.6)
Factual	0.631 (0.775)	N/A	0.798 (0.839)	N/A
Counterfactual	0.822 (0.881)	0.93 (1.222)	1.042 (0.916)	0.449 (0.6)

Table C.13

Summary of readability results for the narratives produced by the system, with “Subject” concerning the instance description only and “Expl.” concerning the explanatory piece only. Reported as the average SMOG Index of the textual piece, with standard deviation in brackets.

<i>Readability</i>	Heart - Subject	Heart - Expl.	Diabetes - Subject	Diabetes - Expl.
Factual + CoT	12.294 (1.661)	13.881 (1.577)	11.694 (1.394)	14.817 (1.586)
Counterfactual + CoT	12.144 (1.798)	15.613 (2.676)	11.611 (1.898)	16.721 (2.456)
Mixed + CoT	11.998 (1.661)	13.601 (1.552)	11.537 (1.638)	14.898 (1.449)
Baseline + CoT	12.326 (1.737)	13.343 (1.169)	10.913 (2.258)	14.949 (1.186)
Factual	12.632 (1.986)	12.927 (1.741)	11.64 (1.29)	14.594 (2.391)
Counterfactual	12.666 (1.625)	15.257 (2.9)	12.096 (1.872)	15.834 (2.645)
Mixed	11.894 (1.645)	13.096 (1.355)	10.977 (1.441)	14.364 (1.683)
Baseline	12.524 (1.207)	14.215 (1.377)	12.68 (1.42)	15.318 (1.243)

C.3. Reported errors

Table C.14

Summary of readability results for the full-text narratives produced by the system, combining description and explanation, reported as the average SMOG Index of the entire piece, with standard deviation in brackets.

Readability	Heart - Full Text	Diabetes - Full Text
Factual + CoT	13.117 (1.1)	13.371 (1.114)
Counterfactual + CoT	13.936 (1.589)	14.446 (1.545)
Mixed + CoT	12.953 (1.175)	13.509 (1.034)
Baseline + CoT	12.969 (0.95)	13.819 (1.073)
Factual	12.723 (1.293)	13.124 (1.379)
Counterfactual	13.881 (1.792)	14.004 (1.633)
Mixed	12.633 (1.001)	12.89 (1.052)
Baseline	13.544 (0.935)	14.456 (0.959)

Table C.15

Summary of the errors occurring during the test runs. “Errors” report instances of the LLM not complying with its instructed format, “Incons.” report instances of the LLM’s prediction disagreeing with its fuzzy-grounded context.

Errors	Heart - Errors	Heart - Incons.	Diabetes - Errors	Diabetes - Incons.
Factual + CoT	15	0	0	0
Counterfactual + CoT	40	991	18	550
Mixed + CoT	1	2	0	3
Baseline + CoT	64	N/A	100	N/A
Factual	0	0	0	0
Counterfactual	1	665	1	357
Mixed	0	0	0	2
Baseline	1	N/A	1	N/A

References

- [1] N. Maslej, L. Fattorini, R. Perrault, Y. Gil, V. Parli, N. Kariuki, E. Capstick, A. Reuel, E. Brynjolfsson, J. Etchemendy, K. Ligett, T. Lyons, J. Manyika, J.C. Niebles, Y. Shoham, R. Wald, T. Walsh, A. Hamrah, L. Santarlasci, J.B. Lotufo, A. Rome, A. Shi, S. Oak, Artificial Intelligence Index Report 2025, 2025, [arXiv:2504.07139](https://arxiv.org/abs/2504.07139) [cs.AI]
- [2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255. <https://doi.org/10.1109/CVPR.2009.5206848>
- [3] L. Nannini, J.M. Alonso-Moral, A. Catala, M. Lama, S. Barro, Operationalizing explainable artificial intelligence in the European union regulatory ecosystem, IEEE Intell. Syst. 39 (4) (2024) 37–48. <https://doi.org/10.1109/MIS.2024.3383155>
- [4] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J.M. Alonso-Moral, R. Confalonieri, R. Guidotti, J.D. Ser, N. Díaz-Rodríguez, F. Herrera, Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence, Inf. Fusion (2023). <https://doi.org/10.1016/j.inffus.2023.101805>
- [5] D. Gunning, E. Vorm, J.Y. Wang, M. Turek, DARPA’s explainable AI (XAI) program: a retrospective, Appl. AI Lett. (2021). <https://doi.org/10.1002/ail.2.61>
- [6] B. Wang, et al., DecodingTrust: a comprehensive assessment of trustworthiness in GPT models, 2024, [arXiv:2306.11698](https://arxiv.org/abs/2306.11698) [cs.CL]
- [7] L.A. Zadeh, Fuzzy sets, Inf. Control 8 (3) (1965) 338–353. [https://doi.org/10.1016/S0019-9958\(65\)90241-X](https://doi.org/10.1016/S0019-9958(65)90241-X)
- [8] L.A. Zadeh, Fuzzy logic, Computer 21 (4) (1988) 83–93. <https://doi.org/10.1109/2.53>
- [9] D. Dubois, H. Prade, The legacy of 50 years of fuzzy sets: a discussion, Fuzzy Sets Syst. 281 (2015) 21–31. <https://doi.org/10.1016/j.fss.2015.09.004>
- [10] C. Kahraman, B. Öztaysi, S. Çevik Onar, A comprehensive literature review of 50 years of fuzzy set theory, Int. J. Comput. Intell. Syst. 281 (2016) 3–24. <https://doi.org/10.1080/18756891.2016.1180817>
- [11] J.M. Alonso, C. Castiello, L. Magdalena, C. Mencar, Explainable Fuzzy Systems - Paving the Way from Interpretable Fuzzy Systems to Explainable AI Systems, Vol. 970, Springer Nature Switzerland AG, 2021. <https://doi.org/10.1007/978-3-030-71098-9>
- [12] P.M. Perez-Ferreiro, A. Catala, A. Bugarin-Diz, J.M. Alonso-Moral, Virtual assistant supported by fusion of fuzzy rule-based explainers, in: 16th European Symposium on Computational Intelligence and Mathematics (ESCIIM), 2024, pp. 1–6.
- [13] 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), 2025, pp. 1–6. <https://doi.org/10.1109/FUZZ62266.2025.11152139>
- [14] N.J. Roese, Counterfactual thinking, Psychol. Bull. 121 (1) (1997) 133. <https://doi.org/10.1037/0033-2909.121.1.133>
- [15] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, F. Turini, Factual and counterfactual explanations for black box decision making, IEEE Intell. Syst. 34 (6) (2019) 14–23. <https://doi.org/10.1109/MIS.2019.2957223>
- [16] P. Lipton, Contrastive explanation, R. Inst. Philos. Suppl. 27 (1990) 247–266. <https://doi.org/10.1017/S1358246100005130>
- [17] J. Wei, et al., Chain-of-thought prompting elicits reasoning in large language models, Adv. Neural Inf. Process. Syst. 35 (2022) 24824–24837. <https://doi.org/10.5555/3600270.3602070>
- [18] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlöterer, M. van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI, ACM Comput. Surv. (2023). <https://doi.org/10.1145/3583558>
- [19] R. Pecori, P. Ducange, F. Marcelloni, Incremental learning of fuzzy decision trees for streaming data classification, in: 11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT), 2019. <https://doi.org/10.2991/eusflat-19.2019.102>
- [20] N. Idris, M.A. Ismail, Breast cancer disease classification using fuzzy-ID3 algorithm with FUZZYDBD method: automatic fuzzy database definition, PeerJ Comput. Sci. 7 (2021) e427. <https://doi.org/10.7717/peerj-cs.427>

- [21] V. Gupta, H. Gaur, S. Vashishtha, U. Das, V. Singh, J. D. A fuzzy rule-based system with decision tree for breast cancer detection, *IET Image Process.* 17 (2023) 2083–2096. <https://doi.org/10.1049/ipr2.12774>
- [22] P. Ducange, M. Fazzolari, F. Marcelloni, G. Miglionico, F. Ruffini, Fuzzy decision trees for explainable brain tumor classification: a comparative study with deep neural networks and classical binary decision trees, *Inf. Syst. Front.* (2026) 1–29. <https://doi.org/10.1007/s10796-025-10683-2>
- [23] T. Miller, Explanation in artificial intelligence: insights from the social sciences, *Artif. Intell.* 267 (2019) 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- [24] I. Stepin, J.M. Alonso, A. Catala, M. Pereira-Fariña, A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence, *IEEE Access* 9 (2021) 11974–12001. <https://doi.org/10.1109/ACCESS.2021.3051315>
- [25] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking, *Data Min. Knowl. Discov.* 38 (5) (2024) 2770–2824.
- [26] R.M.J. Byrne, Counterfactuals in explainable artificial intelligence (XAI): evidence from human reasoning, in: *International Joint Conference on Artificial Intelligence (IJCAI)*, California, CA, 2019, pp. 6276–6282. <https://doi.org/10.24963/ijcai.2019/876>
- [27] G. Tolomei, F. Silvestri, A. Haines, M. Lalmas, Interpretable predictions of tree-based ensembles via actionable feature tweaking, in: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17*, Association for Computing Machinery, New York, NY, USA, 2017, p. 465–474. <https://doi.org/10.1145/3097983.3098039>
- [28] I. Stepin, J.M. Alonso, A. Catala, M. Pereira-Fariña, Generation and evaluation of factual and counterfactual explanations for decision trees and fuzzy rule-based classifiers, in: *IEEE World Congress on Computational Intelligence (IEEE-WCCI)*, 2020, 9177629. <https://doi.org/10.1109/FUZZ48607.2020.9177629>
- [29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, L. Polosukhin, Attention is all you need, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
- [30] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- [31] A. Radford, et al., Improving Language Understanding by Generative Pre-Training, Technical Report, 2018. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
- [32] X. Zhu, C. Zhang, T. Stafford, N. Collier, A. Vlachos, Conformity in large language models, in: W. Che, J. Nabende, E. Shutova, M.T. Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 3854–3872. <https://doi.org/10.18653/v1/2025.acl-long.195>
- [33] A. Srivastava, et al., Beyond the imitation game: quantifying and extrapolating the capabilities of language models, *Trans. Mach. Learn. Res.* 5 (2023) 1–95.
- [34] J. Wei, et al., Emergent abilities of large language models, *Trans. Mach. Learn. Res.* 8 (2022) 1–30.
- [35] A. Elhady, E. Agirre, M. Artetxe, Emergent abilities of large language models under continued pre-training for language adaptation, in: W. Che, J. Nabende, E. Shutova, M.T. Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 32174–32186. <https://doi.org/10.18653/v1/2025.acl-long.1547>
- [36] T. Brown, et al., Language models are few-shot learners, *Adv. Neural Inf. Process. Syst.* 33 (2020) 1877–1901.
- [37] T. Kojima, S.S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, *Adv. Neural Inf. Process. Syst.* 35 (2022) 22199–22213.
- [38] Z. Zhang, A. Zhang, M. Li, A. Smola, Automatic chain of thought prompting in large language models, in: *International Conference on Learning Representations (ICLR)*, 2023.
- [39] N.F. Rajani, B. McCann, C. Xiong, R. Socher, Explain yourself! leveraging language models for commonsense reasoning, in: A. Korhonen, D. Traum, L. Márquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL*, Florence, Italy, 2019, pp. 4932–4942. <https://doi.org/10.18653/v1/P19-1487>
- [40] Q. Ma, X. Ren, C. Huang, XRec: large language models for explainable recommendation, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), *Findings of the Association for Computational Linguistics: EMNLP, ACL*, Miami, Florida, USA, 2024, pp. 391–402. <https://doi.org/10.18653/v1/2024.findings-emnlp.22>
- [41] Y. Lin, Y. Bian, G. Yu, D. Xue, W. Lu, J. Liu, T. Ruan, CoT-planner: chain-of-thoughts as the content planner for few-shot table-to-text generation reduces the hallucinations from LLMs, in: *International Joint Conference on Neural Networks (IJCNN)*, 2025, pp. 1–8. <https://doi.org/10.1109/IJCNN64981.2025.11228570>
- [42] M. Itoh, I. Kobayashi, Toward fuzzy reasoning with large language models, in: *Joint 13th International Conference on Soft Computing and Intelligent Systems and 25th International Symposium on Advanced Intelligent Systems (SCIS&ISIS)*, 2024, pp. 1–6. <https://doi.org/10.1109/SCISISIS61014.2024.10759982>
- [43] M.L. Bangerter, G. Fenza, D. Furno, M. Gallo, V. Loia, C. Stanzone, I. You, A hybrid framework integrating LLM and ANFIS for explainable fact-checking, *IEEE Trans. Fuzzy Syst.* 33 (12) (2025) 4180–4189. <https://doi.org/10.1109/TFUZZ.2024.3431710>
- [44] Y.O. Gat, N. Calderon, A. Feder, A. Chapanin, A. Sharma, R. Reichart, Faithful explanations of black-box NLP models using LLM-generated counterfactuals, in: *The 12th International Conference on Learning Representations*, 2024.
- [45] A. Fredes, J. Vitria, Using LLMs for explaining sets of counterfactual examples to final users, in: *Proceedings of the 2nd Workshop on Causal Inference and Machine Learning in Practice at the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2024, pp. 1–10. <https://doi.org/10.48550/arXiv.2408.15133>
- [46] R.K. Mothilal, A. Sharma, C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, in: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, New York, NY, USA, 2020, pp. 607–617. <https://doi.org/10.1145/3351095.3372850>
- [47] T. Lombrozo, The structure and function of explanations, *Trends Cogn. Sci.* 10 (10) (2006) 464–470.
- [48] T. Lombrozo, Explanatory preferences shape learning and inference, *Trends Cogn. Sci.* 20 (10) (2016) 748–759.
- [49] T. Lombrozo, Simplicity and probability in causal explanation, *Cogn. Psychol.* 55 (3) (2007) 232–257. <https://doi.org/10.1016/j.cogpsych.2006.09.006>
- [50] I. Lage, E. Chen, J. He, M. Narayanan, S. Gershman, B. Kim, F. Doshi-Velez, An evaluation of the human-interpretability of explanation, in: *Conference on Neural Information Processing Systems (NeurIPS) Workshop on Correcting and Critiquing Trends in Machine Learning*, 2018.
- [51] M. Narayanan, E. Chen, J. He, B. Kim, S. Gershman, F. Doshi-Velez, How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation, (2018). <https://arxiv.org/abs/1802.00682>
- [52] E. Dale, J.S. Chall, The concept of readability, *Elem. Engl.* 26 (1) (1949) 19–26. <https://www.jstor.org/stable/41383594>
- [53] R. Flesch, Marks of Readable Style; A Study in Adult Education, *Teachers College Contributions to Education* (1943).
- [54] J.P. Kincaid, R.P. Fishburne, Jr., R.L. Rogers, B.S. Chissom, Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel, Technical Report, 1975. <https://apps.dtic.mil/sti/pdfs/ADA006655.pdf>
- [55] P. Ley, T. Florio, The use of readability formulas in health care, *Psychol. Health Med.* 1 (1) (1996) 7–28. <https://doi.org/10.1080/13548509608400003>
- [56] J.R. Quinlan, Induction of decision trees, *Mach. Learn.* 1 (1986) 81–106. <https://doi.org/10.1007/BF00116251>
- [57] L. Breiman, Random forest, *Mach. Learn.* 45 (1) (2001) 5–32. <https://doi.org/10.1023/A:1010933404324>
- [58] D.P. Pancho, J.M. Alonso, L. Magdalena, Quest for interpretability-accuracy trade-off supported by fingrams into the fuzzy modeling tool GUAJE, *Int. J. Comput. Intell. Syst.* 6 (2013) 46–60. <https://doi.org/10.1080/18756891.2013.818189>
- [59] I. Stepin, J.M. Alonso, A. Catala, M. Pereira-Fariña, An empirical study on how humans appreciate automated counterfactual explanations which embrace imprecise information, *Inf. Sci.* 618 (2022) 379–399. <https://doi.org/10.1016/j.ins.2022.10.098>
- [60] J.M. Alonso, A. Bugarín, ExpliClas: automatic generation of explanations in natural language for weka classifiers, in: *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2019, pp. 1–6. <https://doi.org/10.1109/FUZZ-IEEE.2019.88599018>
- [61] E.H. Mamdani, S. Assilian, An experiment in linguistic synthesis with a fuzzy logic controller, *Int. J. Man-Mach. Stud.* 7 (1975) 1–13. [https://doi.org/10.1016/S0020-7373\(75\)80002-2](https://doi.org/10.1016/S0020-7373(75)80002-2)
- [62] L.A. Zadeh, The concept of a linguistic variable and its application to approximate reasoning—I, *Inf. Sci.* 8 (3) (1975) 199–249. [https://doi.org/10.1016/0020-0255\(75\)90036-5](https://doi.org/10.1016/0020-0255(75)90036-5)

- [63] L.A. Zadeh, Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, *Fuzzy Sets Syst.* 90 (2) (1997) 111–127. [https://doi.org/10.1016/S0165-0114\(97\)00077-8](https://doi.org/10.1016/S0165-0114(97)00077-8)
- [64] E. Reiter, R. Dale, *Building Natural Language Generation Systems*, Cambridge University Press, 2000. <https://doi.org/10.1017/CBO9780511519857>
- [65] E. Reiter, *Natural Language Generation*, Springer Cham, 2024. <https://doi.org/10.1007/978-3-031-68582-8>
- [66] R.R. Hoffman, S.T. Mueller, G. Klein, J. Litman, Measures for explainable AI: explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance, *Front. Comput. Sci.* Volume 5 - 2023 (2023). <https://doi.org/10.3389/fcomp.2023.1096257>
- [67] L. Zheng, et al., Judging LLM-as-a-judge with MT-bench and chatbot arena, *Adv. Neural Inf. Process. Syst.* 36 (2023) 46595–46623. <https://doi.org/10.5555/3666122.3668142>
- [68] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y.J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, *ACM Comput. Surv.* 55 (12) (2023) 1–38. <https://doi.org/10.1145/3571730>
- [69] L.-W. Wang, M.J. Miller, M.R. Schmitt, F.K. Wen, Assessing readability formula differences with written health information materials: application, results, and recommendations, *Res. Soc. Adm. Pharm.* 9 (5) (2013) 503–516. <https://doi.org/10.1016/j.sapharm.2012.05.009>
- [70] H.P. Grice, *Logic and conversation*, in: *Speech Acts*, Brill, 1975, pp. 41–58.
- [71] D. Hilton, Conversational processes and causal explanation, *Psychol. Bull.* 107 (1990) 65–81. <https://doi.org/10.1037/0033-2909.107.1.65>
- [72] J. Li, X. Cheng, X. Zhao, J.-Y. Nie, J.-R. Wen, HaluEval: a large-scale hallucination evaluation benchmark for large language models, in: H. Bouamor, J. Pino, K. Bali (Eds.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Singapore, 2023, pp. 6449–6464. <https://doi.org/10.18653/v1/2023.emnlp-main.397>
- [73] K. Stechly, K. Valmееkam, S. Kambhampati, Chain of thoughtlessness? An analysis of CoT in planning, in: *Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS '24*, Curran Associates Inc., Red Hook, NY, USA, 2024. <https://neurips.cc/virtual/2024/poster/93898>.
- [74] P. Shojaei, I. Mirzadeh, K. Alizadeh, M. Horton, S. Bengio, M. Farajtabar, The illusion of thinking: understanding the strengths and limitations of reasoning models via the lens of problem complexity, 2025, [arXiv:2506.06941](https://arxiv.org/abs/2506.06941) [cs.AI]