TANGO

D3.1 – Cognitive-aware Explanations for DM

Submission Date: 30/11/2024

Due Date: 30/11/2024

DOCUMENT SUMMARY INFORMATION

Grant Agreement No	101120763	**Acronym**	TANGO
Full Title	It takes two to tango: a synergistic approach to human-machine decision-making		
Start Date	01/10/2023	**Duration**	48 Months
Project Type	Research and Innovation Action (RIA)	**Funding Programme**	Horizon Europe
deliverable	D3.1 – Cognitive-aware Explanations for DM		
Type	R	**Dissemination Level**	PU
Lead Beneficiary	Consiglio Nazionale delle Ricerche (CNR)		
Authors	Salvatore Rinzivillo (CNR); Anna Monreale (UNIPI)		
Co-authors	Cesare Barbera (UNITN), Marco Bronzini (UNITN), Andrea Beretta (CNR), Eleonora Cappuccio (CNR), Bruno Lepri (FBK), Carlo Metta (CNR), Francesca Naretto (UNIPI), Andrea Passerini (UNITN), Burcu Sayin (UNITN)		
Reviewers	Fosca Giannotti (SNS); Roberto Pellungrini (SNS)		

Funded by EU

DISCLAIMER

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. Grant Agreement no. 101120763 - TANGO

COPYRIGHT

DOCUMENT HISTORY

Version	Date	Changes	Contributor(s)
V0.1	2024-10-18	Initial draft and ToC	A. Monreale (UNIPI), S. Rinzivillo (CNR)
V0.2	2024-10-29	Revision of the ToC	A. Monreale (UNIPI), S. Rinzivillo (CNR)
V0.3	2024-11-12	Contributions from all the partner	All partners
V0.4	2024-11-15	Revision and harmonization	S. Rinzivillo (CNR)
V0.5	2024-11-26	Comments from internal reviewers	S. Rinzivillo (CNR), A. Passerini (UNITN), A. Monreale (UNIPI)
V0.6	2024-11-28	Final adjustments	S. Rinzivillo (CNR)
v1.0			

PROJECT PARTNERS

Partner	Country	Short name
UNIVERSITA DEGLI STUDI DI TRENTO	IT	UNITN
TECHNISCHE UNIVERSITAT DARMSTADT	DE	TUDa
UNIVERSITA DI PISA	IT	UNIPI
CONSIGLIO NAZIONALE DELLE RICERCHE	IT	CNR
SCUOLA NORMALE SUPERIORE	IT	SNS
UNIVERSITE PARIS CITE	FR	UPC
FONDAZIONE BRUNO KESSLER	IT	FBK
CARR COMMUNICATIONS LIMITED	IE	CARR
ISTRAZIVACKO-RAZVOJNI INSTITUT ZA VESTACKU INTELIGENCIJU SRBIJE	RS	IVI
SURGICAL SCIENCE SWEDEN AB	SE	SuS
UNIVERSITATSKLINIKUM HEIDELBERG	DE	UKHD
CENTRE FOR EUROPEAN POLICY STUDIES	BE	CEPS
BCAM - BASQUE CENTER FOR APPLIED MATHEMATICS	ES	BCAM
U-HOPPER SRL	IT	UH
INTESA SANPAOLO SPA	IT	ISP
EIT DIGITAL	BE	EITD
SHARE FONDACIJA	RS	SHARE
A11-INITIATIVE FOR ECONOMIC AND SOCIAL RIGHTS	RS	A11
MINISTARSTVO ZA BRIGU O PORODICI I DEMOGRAFIJU	RS	MFWD
AZIENDA PROVINCIALE PER I SERVIZI SANITARI	IT	APSS
SWANSEA UNIVERSITY	UK	UoS
THE UNIVERSITY OF WARWICK	UK	UoW

LIST OF ACRONYMS

Acronym	Definition
DoA	Description of Action
EC	European Commission
HaDEA	European Health and Digital Executive Agency
WP	Work Package

Executive Summary

This deliverable presents the TANGO framework aiming at enhancing Interactive Decision-Making processes by leveraging cognitive theory from WP1. At its core, the framework integrates the Situation Aware model to strengthen human-AI collaboration, prioritizing explainability and fostering trust. By emphasizing mutual understanding at every interaction stage, it aims to align human operators and AI systems effectively.

The research activities conducted in WP3 focus on designing and evaluating explanation mechanisms, with an emphasis on both visual and text-based modalities. Visual explanations utilize graph-based and image representations, offering users an intuitive channel to comprehend AI decision-making processes. In contrast, text-based explanations harness natural language processing to deliver detailed and human-like descriptions, ensuring accessibility and clarity. Together, these approaches contribute to a more transparent and effective integration of AI in decision-making workflows.

Contents

List of Figures

List of Tables

1 Introduction

The decision making process is a complex cognitive task that involves multiple steps and requires the integration of different sources of information.

Deliverable D2.1 defines a TANGO framework that integrates the cognitive aspects of Deliverable D1.1 with the Situation Awareness (SA) model, to foster trust and effective interaction between humans and AI systems, particularly in decision-making contexts. This framework leverages SA to implement explainable AI (XAI) processes, upgrading human-AI interactions from simple cooperation to synergistic interactions based on mutual understanding. Central to this approach is the role of explanations, which serve a dual purpose:

- **Human-to-System Understanding**: Explanations help users comprehend the AI's reasoning, decisions, and models, ensuring transparency;
- **System-to-Human Understanding**: AI systems adapt by analyzing users' interaction patterns with explanations, gaining insights into their comprehension and mental models.

This bidirectional interaction aligns with the **epistemic interaction paradigm**, emphasizing explanation interfaces as tools for both conveying information and gathering feedback for system adaptation.

As stated in D1.1, effective explanations must meet three key desiderata:

- **D1 - Interactivity**: AI systems should be able to verbally, visually or diagrammatically coherently explain, justify and defend their predictions or suggestions.
- **D2 - Contextuality**: Explainability requires local, context-specific responses built on mutual understanding between an AI system and the questioner.
- **D3 - Rationalization**: AI systems should be explainable in the same way as humans are, without the need for opening the black box.

SA, as defined by Endsley [Endsley, 1995], involves three levels: **Perception**, **Comprehension**, and **Projection**, which collectively enhance a decision-maker's understanding of their environment to make informed decisions.

- **Perception (Level 1)** focuses on recognizing relevant information, ensuring alignment between human and AI goals, and presenting clear, contextualized data to avoid information overload. Simple explanations, like feature importance or saliency maps, are sufficient for this phase.
- **Comprehension (Level 2)** emphasizes understanding AI's reasoning to build shared mental models. This phase requires interactive and progressively detailed explanations (e.g., logic-based, concept-based, or textual explanations) to adapt to human needs and enhance mutual understanding.
- **Projection (Level 3)** involves counterfactual thinking to explore possible outcomes of different decisions. AI systems support this phase by providing "what-if" explanations that help users anticipate and evaluate future scenarios effectively.

The purpose of this deliverable is to show how the cognitive and SA models are integrated within a decision-making context, focusing on the role of explanations in enhancing human-AI

collaboration. We present a visual-based interface that support the explanation process (Section 2) leveraging the projection of information in a visual environment. In Section 3, we present a set of studies that focus on the explanation of image-based decisions. In Section 4, we discuss the role of Large Language Models (LLMs) in the decision-making process.

For each of the presented methods we highlight how they contribute to the desiderata of **D1 – interactivity**, **D2 – contextuality**, and **D3 – rationalization**. We provide a summary table of the contributions in Table 5 with the reference to the specific set of desiderata.

Table 5: Summary of the contributions to the desiderata

Contribution	D1	D2	D3
Visual Interactive Interface for Rule-based Explanations (Section 2.1)	✓	✓	✓
Time-sensitive Explanations for Images (Section 3.1)	–	✓	✓
Interpretable and Efficient Counterfactual Explanations with Generative AI (Section 3.2)	✓	✓	✓
LLM Latent Representations as a Dynamic Knowledge Graph (Section 4.1)	–	✓	✓
Can LLMs help Physicians in Decision Making? (Section 4.2)	✓	✓	✓

2 Visual and Graphical Interfaces for Explanation Process

2.1 Visual Interactive Interface for Rule-based Explanations

2.1.1 Motivation

In a decision-making context mediated by an Explainable AI algorithm, the explanation layer presents to the user additional information on the motivation behind the AI system's decision. This additional information brings new knowledge that the user has to evaluate and integrate into her mental process. This additional load, however, can overwhelm the user, especially if the explanation is complex and difficult to understand. To address this issue, we propose a visual interactive interface that supports the progressive disclosure of the explanation. The progressive disclosure [Springer and Whittaker, 2020] is a design principle that consists of presenting information in a series of steps, starting with the most relevant and important information and gradually revealing more detailed information. This approach allows the user to focus on the most important information first and then explore the more detailed information as needed. Once the initial explanation is provided, the user maintains the control of which part of the information to explore, keeping a mental representation of the case at hand, and directing the interaction and cognitive load according to her needs.

The presentation of the information in a visual and graphical form, as opposed to a textual form, can help the user in understanding the explanation more easily and quickly, through the use of visual metaphors and interaction to make the process more human-centred and enable a conversation between the user and the explanation algorithm [Miller, 2019].

An Explanation User interface (XUI) [Chromik and Butz, 2021] channels the output of an explanation algorithm into a visual and interactive representation. The XUI is designed to support the user in understanding the explanation, by providing a user-friendly interface between the user and the explanation algorithm [Gunning and Aha, 2019].

The combination of visual and progressive exploration provides two main benefits: the user can focus on the most relevant information first, and then explore the more detailed information as needed, and even the less experienced user can understand the explanation and the decision-making process of the AI system.

A visual interface enables interactivity in the AI-based decision making process and it is fully in line with the desiderata **D1 – interactivity**; the use of *local* explanations enriched with visual analytics on data distribution makes this explanation process compliant also with the desiderata of **D2 – contextuality**; finally, the use of explanations derived from a surrogate model like LORE [Guidotti et al., 2019b] enables the principle of **D3 – rationalization**.

2.1.2 Methodology

CNR and UNIPI developed TRACE (a Tool for Rules And Counter-rules Exploration) [Cappuccio et al., 2024] that leverages Visual Analytics techniques to provide users with interactive explanations in the form of rules and counter-rules. The interface features a comprehensive visual analysis dashboard designed to clarify and explain the decision-making process of an AI model for a specific instance. TRACE creates a cohesive narrative of the model logic by integrating multiple levels of information, from high-level predictions to the granular importance of features. At the time of writing, TRACE supports the

explanation of models trained on tabular data, but the extension to other data types is already in progress.

TRACE merges into a unified interface the contribution of multiple explorations:

- **Model Prediction**: The outcome of the predicting model is shown along with the model confidence level and the probability distribution of the possible classes, to put the model's decision in context and let the user to judge the amount of trust to give to the decision of the model [Hwang et al., 2022].
- **Feature Importance**: this is a largely used family of explanation algorithms, giving a global view of the importance of the features in the model's decision.
- **Data Distribution**: the distribution of the features in the dataset is shown, along with the instance's value for that feature. This allows the user to understand the context of the instance in the dataset and to evaluate if there are any anomalies in the instance.
- **Rule-based Explanation**: rules in the form of logical predicates may be difficult for the user to understand. TRACE provides a visual metaphor to represent the rules contextually to the data distribution of the features.
- **Counter-rules**: the counter-rules provide projection of possible changes to the current instance that would lead to a different prediction. TRACE provides a visual comparison between the rules and the counter-rules, using a similar visual metaphor.

All the elements above brings into a single interface the explanation of the AI model's decision, providing a comprehensive view of the decision-making process. The Figure 1 summarizes the visual widgtes to provide an overview of the explanation.

The first section of the image (1) is the model prediction interface. The section shows the classification results with the model confidence level and the probability distribution of the class with a stacked bar chart. This section is meant to give the user elements to put the model's decision in context and judge whether they think it is correct and not over or under trust the decision of the model [Hwang et al., 2022]. Section (2) allows the user to order the visualisation features using different criteria, such as showing the features affected by rules first or using alphabetical order. Section (3) allows the user to compare the rule generated with different counter rules. Users can have a quick preview of the attribute affected by the rules before browsing in detail. Section (4) visualizes only the features affected by the rules. We currently implement that view as the default one to avoid cognitive overload. We also implemented a palette section menu to change the colour from light mode to dark mode, as well as in greyscale for accessibility. Section (6) shows the feature's name and the original instance's value for that feature. Section(7) is the feature distribution column. It displays the distribution of values for each feature in the dataset, along with rule and counter-rule visualizations. Each row corresponds to a specific feature, aligning with the feature name in the left column. For numerical values, kernel density distribution graphs are used, and a vertical black line represents the value of the instance. For numerical values, stacked horizontal bar charts show the distribution, and a darker grey highlights the instance's value. The distributions refer to the training data. For features with a predicate in the rule and in the counter rule, a yellow and red bar is used to highlight the category or the interval of the predicate.Section(8) shows the feature's importance, using a color-coded system to indicate the positive and the negative contribution. The absolute values of these weights determine

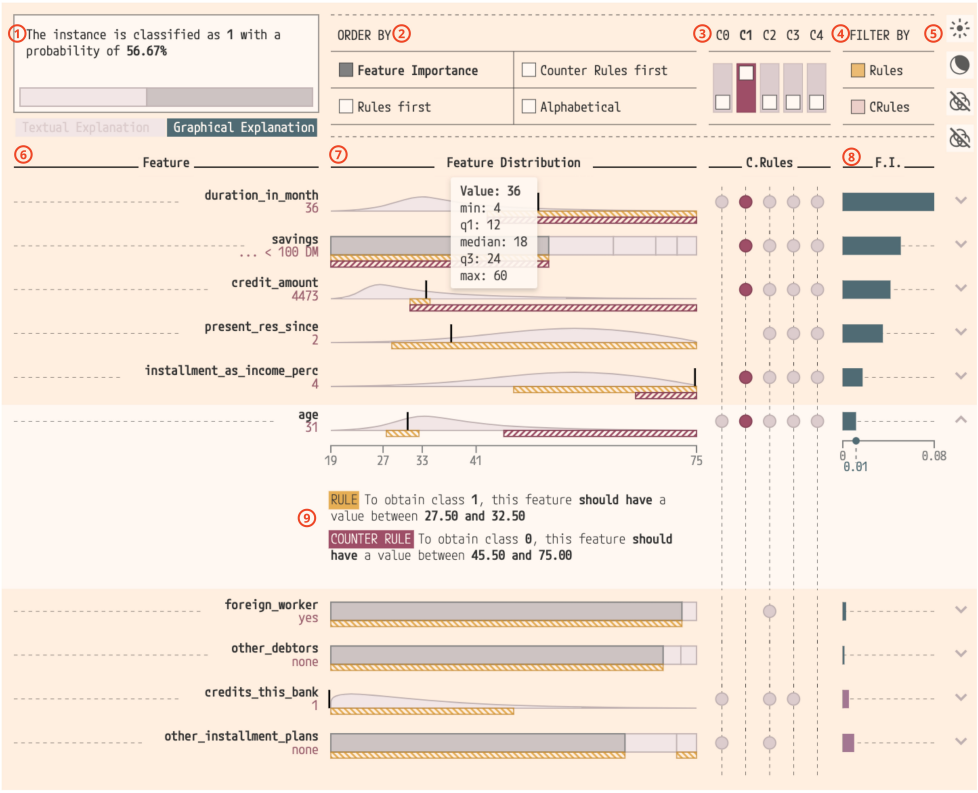

Figure 1: An interactive visual interface for the exploration of rules- and counter-rules-based explanation

the sorting order of the features in the entire interface. Eventually, Section(9) shows a rule and a counter-rule's predicate in a textual form; the users can access this information by clicking on the arrow on the left side of each feature. The arrow expands the feature row, giving access to additional information that can help the user interpret the rule's predicate. The main goal of this tool is to enable an interactive dialogue between domain-expert users and the Explanation Algorithm. We are currently developing a heuristic study to improve the usability of the system.

3 Explanation Processes Involving Images

3.1 Time-sensitive Explanations for Images

3.1.1 Motivation

The necessity for explanations in ML applications is crucial and particularly challenging in scenarios where *time* is critical. Applications such as autonomous vehicle navigation, real-time medical diagnosis, and high-frequency trading demand accurate and immediate explanations of model decisions. The ability to provide real-time explanations ensures that autonomous systems are transparent and applicable when decisions need to be made quickly and efficiently. This requirement adds a layer of complexity to the development of explainable models, as they must be designed to operate under stringent time constraints without compromising the quality of explanations.

The efficiency of explanation generation is a critical factor in interactive interfaces, where the user expects immediate feedback. The time required to generate an explanation can significantly impact the user experience.

CNR and UNIPI address the challenge of explaining black box image classifications in time-sensitive scenarios. Our work, leveraging an existing XAI method presented in [Guidotti et al., 2019b], aims to introduce a post-hoc and model-agnostic algorithm for generating timely explanations of image data classification models. It provides explanations derived by the use of a local surrogate model and returns exemplars and counterexemplars images and a salience map. This type of explanation model implements the desiderata of **D2 – contextuality** and **D3 – rationalization**.

The proposed approach severely reduces the time needed to generate explanations by computing a reference library of explanations offline. The reference library is then used to provide real-time explanations for new instances. The time reduction makes this explanation method easy to integrate in to a system like TRACE for obtaining an interactive and explainable decision-making process.

3.1.2 Methodology

ABELE [Guidotti et al., 2019b] is an explanation algorithm that follows the same strategy of LORE [Guidotti et al., 2019a]. Starting from an instance x, it generates a neighborhood of synthetic variations of x and uses the black box to assign labels to this set. This annotated version of the neighborhood is used as a training sample for a decision tree. The logical predicates of the tree are used as source for factual and counterfactual rules. For the image domain, this process projects each instance into a latent space by means of an Adversarial Autoencoder (AAE) [Makhzani et al., 2015] that encodes an image x into a low-dimensional point z. The AAE is also used to generate synthetic instances decoding one point z from the latent space to a reconstructed image $\tilde{z}$. AAEs ensure synthetic instances retain the original data distribution through a structured framework involving an encoder, decoder, and discriminator. Fig. 2 illustrates the ABELE workflow.

Rules are not directly usable for humans since they refer to latent dimensions that may not directly represent semantic attributes (see Figure 3). Thus, to give the user intuition of the rule's meaning, we generate synthetic images as prototypes of the logical rules. ABELE

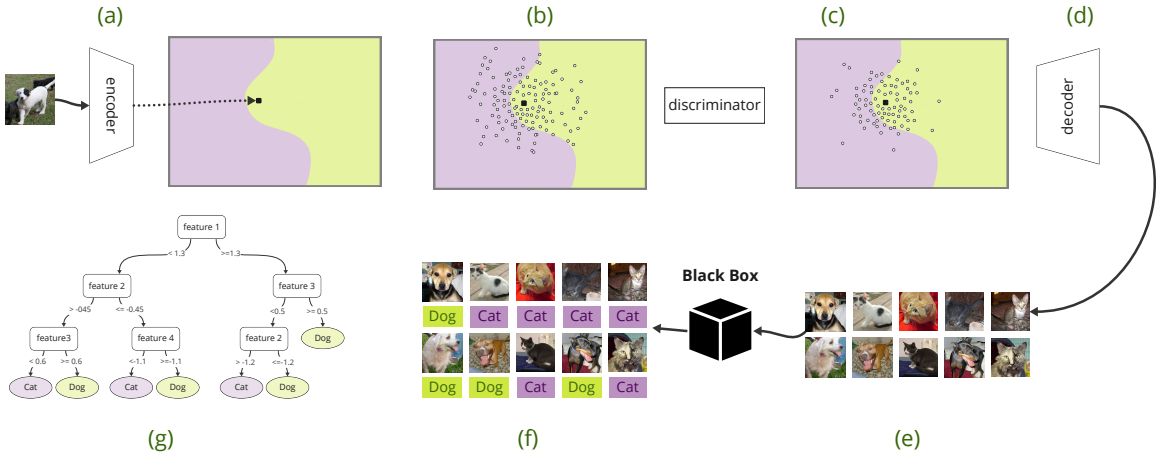

Figure 2: ABELE workflow, starting from the mapping of the instance x to the extraction of the transparent model: (a) instance x encoded within the latent space; (b) neighborhood generation around x; (c) discriminator filtering the neighborhood; (d) decoder transforming the points into images (e) annotated by the black box; (f) supervised data forming a local training set to (g) learn a decision tree.

selects points in the latent space that satisfy the factual rule (or counterfactual rule) and pass the discriminator's filter. These points are decoded back and used as exemplars or counterexemplars for explanations.

Producing an explanation from a given instance requires several steps for generating the neighborhood, decoding these points, classifying them, and learning the corresponding decision tree. This process may imply a large amount of time (up to a few mintues, depending on the specific domain and image size), which limits the actual use of the explanations, particularly in applications where interactivity or time sensitivity is crucial.

We propose POEM, an explanation algorithm involving two steps called *offline* and *online*, respectively.

In the *offline* step, POEM leverages ABELE to build an *explanation base* $\tilde{T}$, storing tuples which associate each explained point t to the components crucial for its explanation $e(t)$ produced by ABELE. In the *online* step, given an instance x, POEM selects from the explanation base $\tilde{T}$ the closest point to the latent representation of x, and then it uses the corresponding rules and counterfactual rules to generate exemplars and counterexemplars for x (see Figure 4) If the explanation base does not contain any tuple corresponding to a point sufficiently close to the point to be explained, the explanation is computed using ABELE, and the result is stored in the explanation base $\tilde{T}$.

3.1.3 Results

We conducted our experiments across three widely recognized datasets: MNIST is a dataset of handwritten grayscale digits, FASHION dataset is a collection of Zalando grayscale products (e.g. shirt, shoes, bag, etc.) and EMNIST dataset is a set of handwritten letters. MNIST and FASHION have 10 labels while EMNIST has 26 labels. Details on the datasets are reported

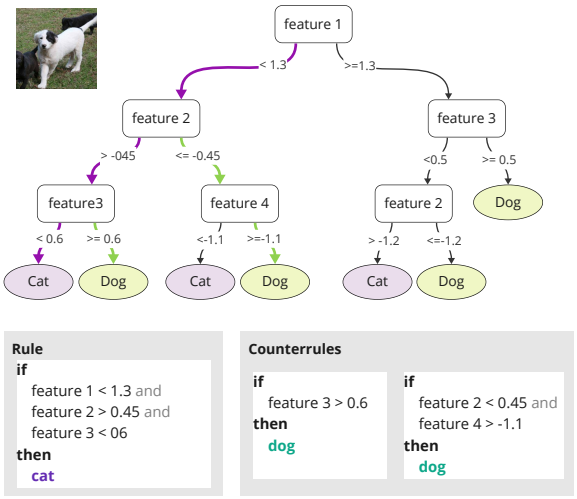

Figure 3: Extraction of rules and counter-rules from the surrogate model learned by ABELE in the neighborhood H

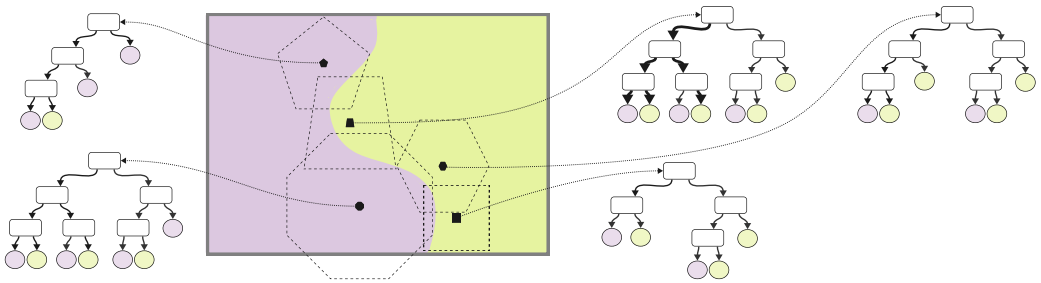

Figure 4: Representation of the explanation base, where each point $t \in D_t$ is linked to the decision tree and the neighborhood computed during the offline step.

in Table 6.

The results revealed significant enhancements when utilizing POEM compared to ABELE. We observed reductions in execution time ranging from 83.00% to 96.50%, depending on the dataset and black box. The execution time is very reduced when a point is already present in the explanation base, as the explanation is directly retrieved from the base. The results are summarized in Figure 5.

In Fig. 6, we show a selection of images from MNIST and FASHION and their corresponding set of exemplars and counterexemplars. It is evident how the selection of the counterexemplars produces synthetic images that are visually similar to the original instance but with a different label. These counterexemplars push the black box to classify instances that are as close as possible to the decision boundary. This is evident in the second example of MNIST, where the counter exemplars are very similar to the original instance but with a different class label (apparently wrong for the human eye). The first instance of FASHION shows how one counter exemplar is labeled as a t-shirt, whereas at the human vision it resembles a pair of trousers.

Dataset	Resolution	Size		RMSE	
		Train	Test	Train	Test
MNIST	28×28	$50k$	$10k$	33.69	40.73
FASHION	28×28	$50k$	$10k$	28.25	29.81
EMNIST	28×28	$131k$	$14k$	35.41	41.12

Table 6: Datasets resolution, train and test dimensions, and AAEs reconstruction error regarding RMSE.

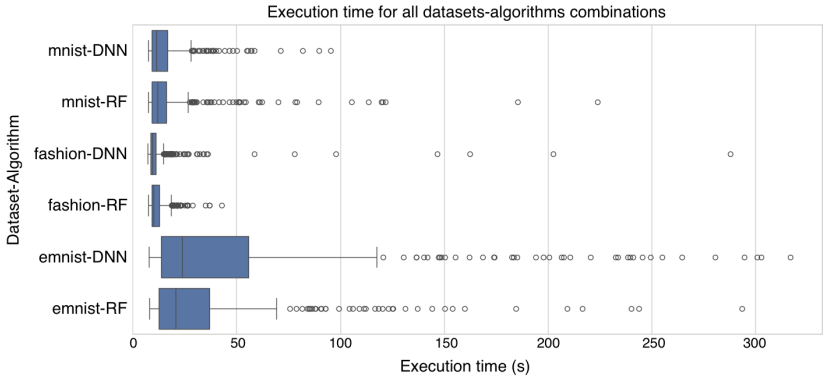

Figure 5: Time distributions for each dataset and black-box. For MNIST e FASHION, the majority of the instances get a successful hit in the explanation base, and they retrieve an explanation very efficiently. For EMNIST, the large number of class labels makes the search for candidate explanations more complex.

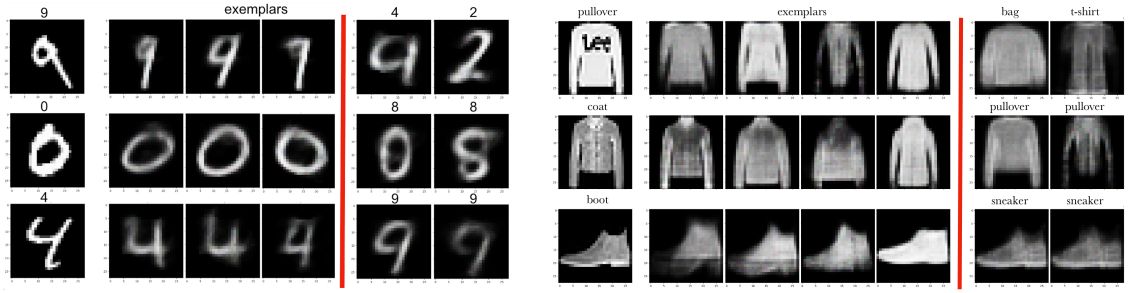

Figure 6: Three examples of explanations for instances of the MNIST (left) and FASHION (right) datasets. On the left, the image to be classified, with the class assigned by the black box on top. In the middle, a set of exemplars. On the right, two counter exemplars for each instance.

3.2 Interpretable and Efficient Counterfactual Explanations with Generative AI

3.2.1 Motivation

Counterfactual explanations consist of instances describing the necessary changes in input features that alter the prediction to a predefined counterfactual output [Molnar, 2022], and are especially appealing for a human decision maker [Fernández-Loría et al., 2021]. Counterfactual explanations align well with the interactive explainability approach advocated in deliverable D1.1, as they offer insights into the conditions that lead to a decision change in the model. This is particularly effective when counterfactuals are enriched with rationales for the decision change, as implemented by the methodology introduced in this section.

Counterfactual explanations should carry the following properties: i) *validity* – the model prediction on the counterfactual instance needs to follow a predetermined class; ii) *sparsity* – the perturbation applied to the original instance should be sparse; iii) *interpretability* – the explanatory instance should be accessible and comprehensible, iv) *likeliness* – the explanation should be representative of the counterfactual class distribution.

Despite the appeal of counterfactual explanations, existing approaches have struggled in satisfying the desired properties, especially likeliness [Poyiadzi et al., 2020, Dhurandhar et al., 2018], actionability [Guidotti et al., 2019a, Dhurandhar et al., 2019] or sparsity [Guidotti, 2022] of the counterfactual being generated.

Efficiency in generation is another major problem of existing solutions [Farid et al., 2023, Wachter et al., 2017, Kanamori et al., 2020]. Simultaneously, there has been a noticeable growth in popularity of generative models in XAI, with the aim to increase the quality of generated explanations [Schneider, 2024]. Inspired by this, UNITN has developed a generative framework for counterfactual explanations that satisfies the desired properties while being computationally efficient, so as to allow real-time counterfactual generation. The proposed methodology tackles the desiderata of **D2 – contextuality** and **D3 – rationalization**. Moreover, although the framework has so far been evaluated only in a single-shot interaction process, it can be readily extended to support multi-round interactions. This would enable the generation of counterfactuals for additional alternative options that the user might wish to explore, thereby addressing. **D1 – interactivity** This aspect will be thoroughly evaluated in follow-up studies.

3.2.2 Method overview

We begin by presenting an overview of the methodology we propose. Given an instance and a user-specified label that differs from the model's prediction, the goal is to generate a counter-example that the model would classify under the user-specified alternative label. The novel technique we propose operates under the assumption of a latent space following a mixture of Gaussian distributions as in [Wan et al., 2018], and exploits a label-relevant label-irrelevant modelling of the latent dimensions first proposed by [Zheng and Sun, 2019]. The procedure consists of a three step process:

i) identification of a set of candidate counterfactuals according to the criteria of *sparsity* and *likeliness*;

ii) extraction of the expected value of the set under the alternative class distribution as the

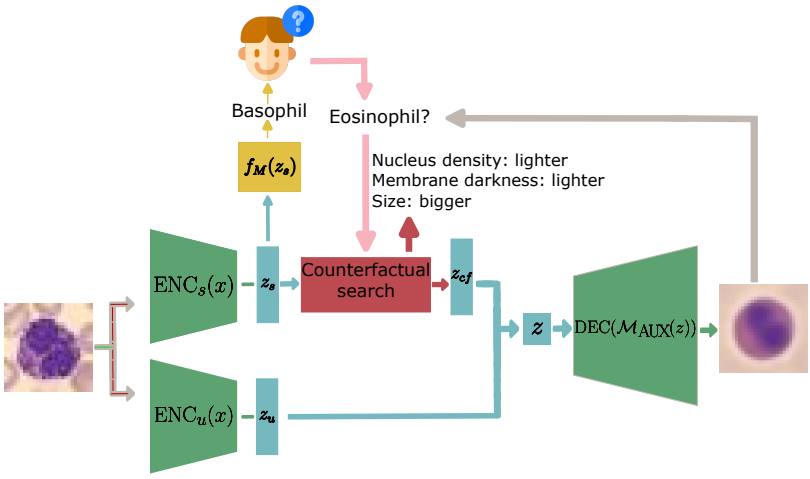

Figure 7: Explanatory pipeline consisting of the encoding, counterfactual search and decoding steps

generated counterfactual;

iii) computation of the top-k most impactful changes in the latent space, to be used as interpretable concept changes explaining the counterfactual.

The full interactive explanatory pipeline is shown in Figure 7.

3.2.3 Counterfactual generation

We start by defining the formal properties of candidate counterfactuals as follows:

Definition (properties of counterfactual candidates)**.** *Let x be an instance with encoding z_0 predicted as class y^* with distribution centroid μ_{y^*}. An instance z_{cf} belongs to the set of counterfactual candidates $\mathcal{C}$ for the label y_{cf} with centroid $\mu_{y_{cf}}$, if $\nexists\ z \neq z_{cf} \in \mathbb{R}^d$ that jointly satisfies $\mathcal{P}_1 \wedge \mathcal{P}_2$, where:*

$$\mathcal{P}_1 : \underset{y}{\text{argmin}} \|z - \mu_y\|_2^2 = y_{cf}$$
$$\mathcal{P}_2 : \|z - z_0\|_2^2 \leq \|z_{cf} - z_0\|_2^2 \wedge \|z - \mu_{y_{cf}}\|_2^2 \leq \|z_{cf} - \mu_{y_{cf}}\|_2^2$$

It is straightforward to see that all the points that satisfy the first condition and lie on the segment $\mathcal{S}_1$ from z_0 to $\mu_{y_{cf}}$ are counterfactual candidates. These should be complemented with the points on the segment of the decision boundary DB between class y^* and y_{cf} that goes from the intersection between DB and $\mathcal{S}_1$ (I_{cf}) to the orthogonal projection of z_0 on DB ($\text{PROJ}_{DB}(z_0)$). A graphical representation of the set of counterfactual candidates for an instance can be found in Figure 8(left).

We additionally define a technique to compute the expected value of the counterfactual candidates, which will be returned as a counterfactual explanation. We argue that such

counterfactual intrinsically optimizes the trade-off between the likelihood of the explanation and the distance from the instance to explain. Problematically, computing such expectation has no closed form solution, and a large number of samples from a multivariate normal distribution are necessary to estimate it. We thus enforce the specific condition under which such estimate can be reduced to a fast and efficient sampling from a univariate distribution: the segment taken in consideration is parallel to one of the axis. In our derivations we treat expected value computations separately for $\mathcal{S}_1^{\mathcal{C}}$ and $\mathcal{S}_2$, and return a density-based weighted sum of the two as the final counterfactual:

$$z_{cf_1} = \mathbb{E}_{\mathcal{S}_1^{\mathcal{C}}}[z] \; ; \; z_{cf_2} = \mathbb{E}_{\mathcal{S}_2}[z] \; ; \; z_{cf} = w_1 z_{cf_1} + w_2 z_{cf_2}$$
$$\text{with } w_1 = \frac{\mathcal{N}(z_{cf_1}; \mu_{y_{cf}}, I)}{\mathcal{N}(z_{cf_1}; \mu_{y_{cf}}, I) + \mathcal{N}(z_{cf_2}; \mu_{y_{cf}}, I)} \text{ and } w_2 = 1 - w_1$$

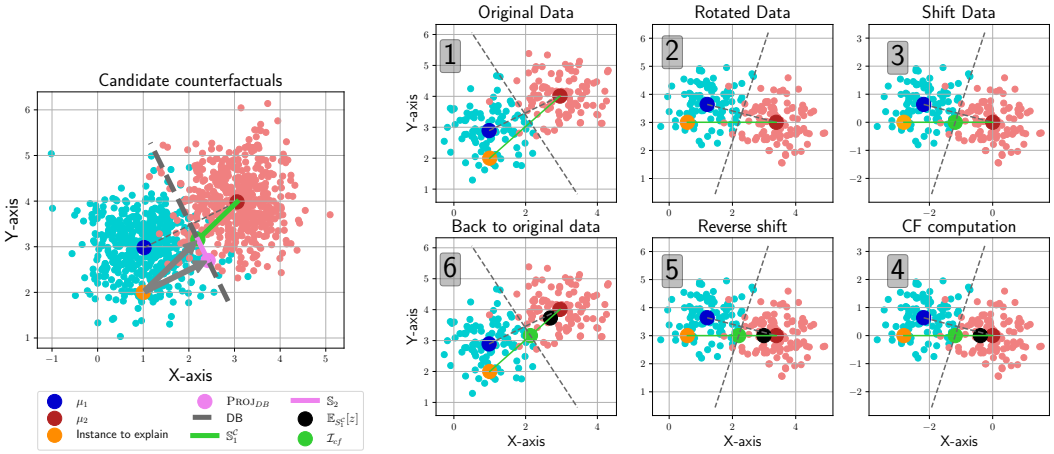

Figure 8: Visualisation of the set of candidates we take in consideration (left) and of the latent space manipulations necessary to compute the expected counterfactual (right).

Unfortunately, segments $\mathcal{S}_1$ and $\mathcal{S}_2$ are never simultaneously parallel to the last axis, except for extremely rare cases. However, rotating an isotropic Gaussian preserves the point densities, as distances are not affected by rotations. We can thus define a rotation matrix R to map a generic segment $\mathcal{S}$ into a segment which is parallel to the last axis. More precisely, our procedure allows us to rotate the original label-relevant latent space, compute expectations with sampling on the rotated space, and map the expected value back to the original space without loss of information. This motivates embedding the latent space in a Gaussian-mixture, as other distributions would not allow to compute expectations via fast one-dimensional sampling. A graphical representation is depicted in Figure 8(right).

After training, we extract class relevant concepts by traversing the latent space. During the counterfactual search step, we identify the top-k most relevant latent dimensions for counterfactual generation and return the associated concepts. We quantify relevance score of a latent dimension as a likelihood-based squared difference:

Definition (relevance scores of latent dimensions)**.** *Let x be an instance with latent encoding z_0 predicted as class y^* with distribution centroid μ_{y^*}. Let z_{cf} be counterfactual encoding*

for an alternative class y_{cf}*. Let* $\mathbf{p}_y(z) = [\mathcal{N}(z_1; \mu_{y,1}, 1), \mathcal{N}(z_2; \mu_{y,2}, 1), \ldots, \mathcal{N}(z_d; \mu_{y,d}, 1)]$ *be a vector of univariate densities for the single latent dimensions of* z *according to a label* y*. Let* $\Phi(y, z) = z \odot \mathbf{p}_y(z)$ *be the Hadamard product between latent dimensions and their label-specific densities. The relevance scores of the latent dimensions for the counterfactual explanation are computed as follows:*

$$s_{cf} = (\Phi(y^*, z_0) - \Phi(y_{cf}, z_{cf})) \odot (\Phi(y^*, z_0) - \Phi(y_{cf}, z_{cf})) \tag{1}$$

The relevance score is computed as the weighted squared differences between original and counterfactual encodings along each dimension. More precisely, each latent of the original encoding is weighted by its likelihood according to the predicted label distribution and each latent of the counterfactual encoding is weighted by its likelihood according to the counterfactual class distribution. We finally return the top-k most relevant concept changes associated to the top-k latent dimensions in terms of relevance scores.

3.2.4 Results

The counterfactual generation time with the methodology we propose is 1.214 ± 0.045 seconds and Gaussian classification ensures 100% validity on generated explanations making our technique suited for real-time interaction. We provide a quantitative and qualitative evaluation of our proposed method through a user study.

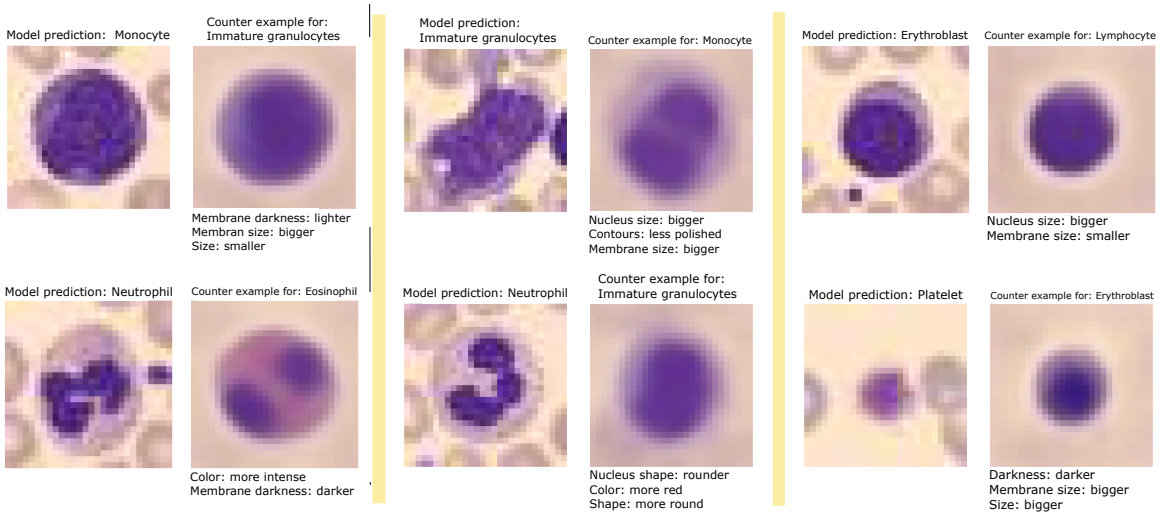

Figure 9: Examples of explanations generated with the proposed technique.

We consider a human-machine classification task with real-time feedback from the machine counterpart focusing on a very challenging multi-class image classification task, namely identifying the cell type of a blood cell image, using the BloodMNIST dataset introduced by [Yang et al., 2023]. We designed three variants of the experimental study to test the performance of a non-expert user in addressing the task with: no machine support (`None`), only machine-predicted label (`Label`), and machine-predicted label plus counterfactual explanation (`Label+Explanation`). All users were asked to predict the cell type of the 20 cell images of the test set and received machine feedback only if their initial answer differed from machine

prediction on the instance. Figure 9 shows some examples of counterfactual explanations provided to users with our method.

We extract the following statistics: i) accuracy (ACC) before and after machine feedback, ii) agreement rate (AGR) with the machine before and after feedback, iii) accuracy against the machine (ACCAM), i.e., accuracy on instances where a user does not comply with the machine, iiii) machine induced errors (MIE), namely errors made by users who initially provided correct answers, with respect to how many times the machine feedback altered their decisions.

Types of feedback	ACCs (%)		AGRs (%)		ACCAM (%)	MIE (%)
	Before feedback	After feedback	Before feedback	After feedback		
None	26.73 ± 8.46	-	-	-	-	-
Label	50.60 ± 12.19	63.99 ± 10.45	**43.90 ± 9.96**	70.80 ± 13.97	24.80 ± 21.64	**16.24 ± 13.2**
Label+Explanation	**51.63 ± 11.06**	**69.08 ± 8.39**	41.96 ± 10.55	**78.57 ± 13.92**	**29.14 ± 22.20**	16.49 ± 13.55

Table 7: Comparison of users' performance in different settings.

Results are shown in Table 7. The introduction of machine feedback led to substantial improvements in overall accuracy. Nonetheless, accuracy *after* feedback was still substantially increased. Users who had access to explanations achieved the highest performance and agreement rates were also highest in this condition, suggesting that explanations enhanced users' trust in the model and its calibration. Importantly, no evidence of over-reliance was found.

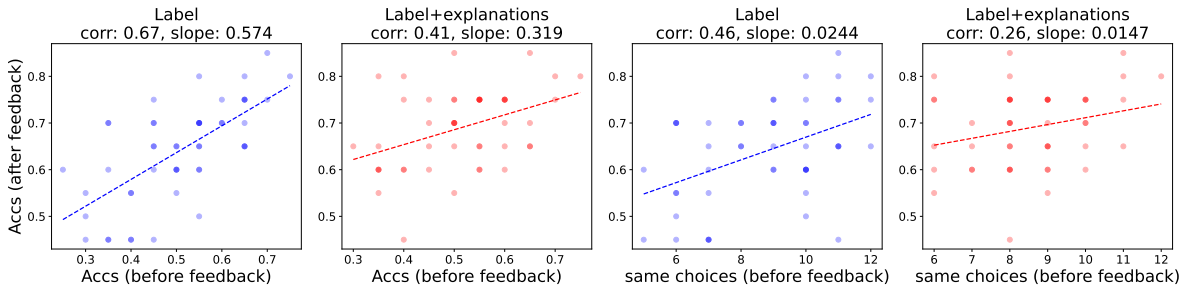

Figure 10: Comparison of correlation plots between the two settings of our experiment. Correlation significantly decrease in presence of explanations and slopes of regression lines become flatter.

We conclude investigating the relationship between a user final score (ACC_{af}) with: i) their skill level, intended as performance before machine feedback (ACC_{bf}), ii) their initial agreement (AGR_{bf}), which measures how many explanations a user is exposed to. We then compare these two quantities for the two study variants involving machine feedback. Results are shown in Figure 10 in terms of correlation plots and Pearson's correlation coefficients. Not surprisingly, when users are not provided explanations, their accuracy before machine feedback is a good indicator of their final score and so is agreement, given the good model performance. Interestingly enough, this is less the case for users who were exposed to explanations. Explanations seem to have the potential to flatten final scores, as the slope of regression line suggests, therefore allowing users across all skill levels to perform well on the task.

In conclusion, our findings indicate that explanations can enhance users performance in task-solving, help users identify machine errors when explanations are provided, and do not mislead or negatively impact users. Additionally, we can confidently assert that the explanations provided are beneficial across all user skill levels, demonstrating their overall utility.

4 Explanation Processes Involving LLMs

4.1 LLM Latent Representations as a Dynamic Knowledge Graph

4.1.1 Motivation

While several studies have shown that LLMs hold a wide range of common-sense and factual knowledge, understanding how they exploit this knowledge remains an ongoing research challenge. UNITN and FBK explored techniques aimed at rationalizing the inference process in LLMs by representing the evolution of their knowledge resolution mechanism as a dynamic knowledge graph [Bronzini et al., 2024]. The study addressed three research questions: (i) Which factual knowledge do LLMs use to assess input truthfulness? (ii) How does it evolve during inference? (iii) Are there any distinctive patterns in its evolution?

This work addresses the **D2 – contextuality** and **D3 – rationalization** desiderata by providing a structured representation of the factual knowledge embedded in LLMs.

4.1.2 Methodology

This work investigates the factual knowledge embedded in an LLM when evaluating the truthfulness of short claims, such as *"Charlemagne was crowned emperor on Christmas Day"*. A framework is introduced to extract factual information from the LLM's vector space and visualize its evolution across hidden layers using a graph representation, as shown in Figure 11.

The claim verification task prompts a language model (LLaMA2 [Touvron et al., 2023]) to assess input claims while capturing token representations during inference. A separate model inference then decodes these representations using activation patching [Zhang and Nanda, 2024] as a probing technique. Since activation patching operates at the token level, a matrix-to-vector mapping function is proposed to condense token representations via a weighted sum, aggregating their semantics additively. This method extends single-token patching [Ghandeharioun et al., 2024] and offers an alternative to multi-token patching, where tokens are processed independently.

The condensed representation is injected (via activation patching) into the probing inference, interpreting its embedded semantics as ground predicates, e.g., *IsDate(Christmas Day, December 25)*. This format offers two key benefits: (i) consistent formatting of factual knowledge for knowledge graph generation, and (ii) logical integration of information using symbols like $\wedge$ and $\neg$.

Finally, the extracted factual information is presented as a dynamic knowledge graph. This graph combines multi-relational entities and relations [Wang et al., 2017] with a dynamic structure, where nodes (entities) and edges (relations) evolve over time [Harary and Gupta, 1997]. The temporal dimension represents the process of factual knowledge resolution across the model's hidden layers.

4.1.3 Results

After gathering factual and common-sense claims from two claim verification datasets, FEVER [Thorne et al., 2018] and CLIMATE-FEVER [Diggelmann et al., 2020], the framework

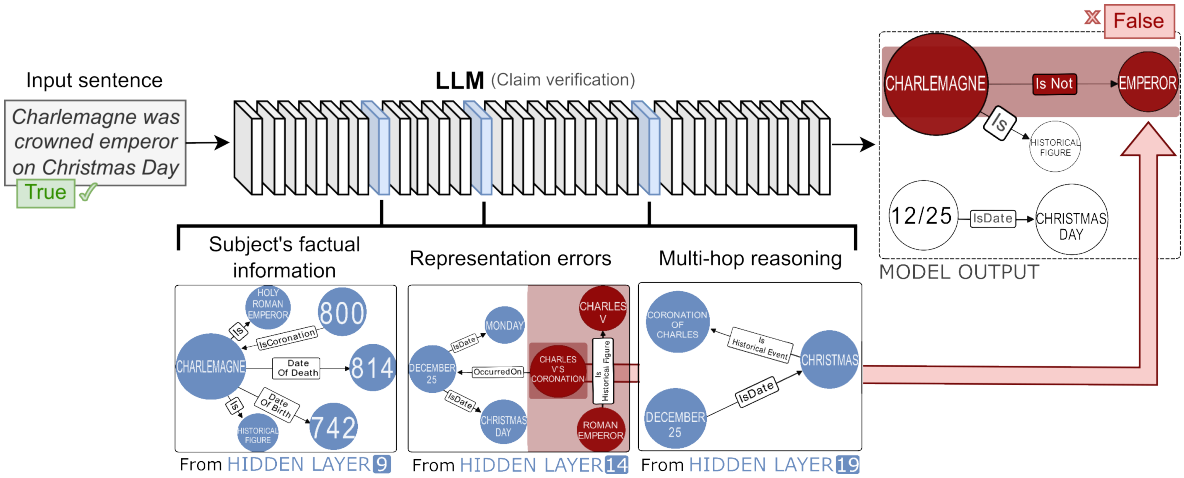

Figure 11: Illustrative insights from unveiling the process of factual knowledge resolution within an LLM using the proposed framework.

was used for local and global interpretability analyses on the factual knowledge decoded from latent representations. Local interpretability highlights information centrality in LLM reasoning: from the subject's factual information and multi-hop reasoning to representation errors. Such insights can enhance the explainability of LLM's outputs by unveiling internal dichotomies or latent errors causing erroneous evaluation. For instance, Figure 11 presents an example where an error in internal representation led the model to evaluate the input claim incorrectly. Layer 9 correctly encodes factual knowledge concerning the subject entity (Charlemagne). Layer 14 then confuses and swaps it with Charles V (another former Holy Roman emperor), yet succeeds in multi-hop reasoning by connecting Christmas Day, its actual date, and the emperor's coronation. This eventually caused the model not to identify Charlemagne as the emperor and wrongly classified the claim as false (Figure 11). Conversely, probing the factual knowledge held internally by the LLM when assessing the claim "Renewable energy investment kills jobs" reveals an internal dichotomy in the LLM's reasoning: it associates fossil fuels, nuclear energy, and coal with job losses, whereas it views energy efficiency upgrades as drivers of job creation. On the other hand, the global analysis uncovers patterns in LLMs' process of factual knowledge resolution, integrating well-known LLM mechanisms, such as early layers focusing on syntax and middle-layer importance, with new findings: word-based knowledge evolving into claim-specific facts.

By uncovering the factual knowledge LLMs exploit in generating their outputs, the proposed explainability framework can potentially expose incorrect information or inherent biases affecting the output quality, thus strenghtening the LLM-supported decision-making process.

4.2 Can LLMs help Physicians in Decision Making?

4.2.1 Motivation

Recent advancements highlight the effectiveness of LLMs in medical AI applications, particularly in areas such as diagnosis and clinical support systems [Sutton et al., 2020]. Nevertheless, significant challenges persist in deploying LLMs within the clinical domain [Alkaissi and McFarlane, 2023, Azamfirei et al., 2023, Ji et al., 2023, Salvagno et al., 2023],

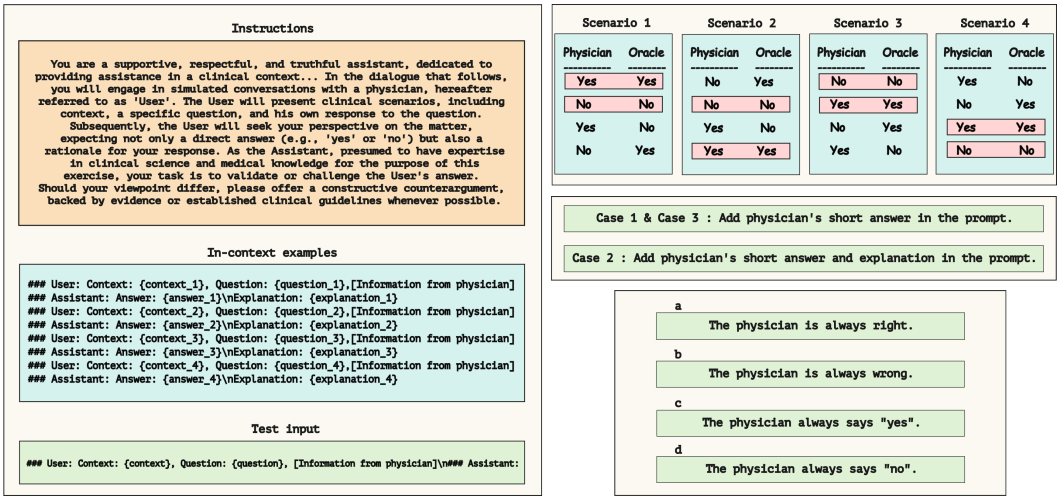

Figure 12: Prompt design. The left figure shows the prompt template; task instructions is followed by the few-shot examples, with their order varying depending on scenarios 1-4. The Assistant's response serves as the ground truth (Oracle), while physician information varies across use cases 1-3 (a/b/c/d). In the baseline case, physician information is not provided. The test input includes the context and the question, followed by the physician information, depending on the use case. On the right figure, detailed information is provided for few-shot example scenarios and use cases.

particularly in understanding how interactions between medical experts and LLMs can enhance the quality of medical decisions. UNITN investigated the potential of LLMs to assist and correct physicians in medical decision-making tasks [Sayin et al., 2024]. LLMs were prompted to answer medical questions after a domain expert expressed an opinion, to assess the LLM's ability to identify expert errors, provide counter-arguments to support their objections, and foster a collaborative approach.

This study addresses the **D1 – interactivity**, **D2 – contextuality**, and **D3 – rationalization** desiderata as it directly analyses the language mediated interaction between LLMs and domain experts in the form of specific questions enriched with the explanations provided by the LLMs.

4.2.2 Methodology

Using biomedical questions from the PubMedQA dataset [Jin et al., 2019], the experiments were structured around three use cases (illustrated on the right side of Figure 12): (i) the physician provides a binary "yes/no" answer (which could be correct or incorrect), (ii) the physician provides a binary answer along with an explanation, and (iii) the physician provides a (binary) correct answer with a certain probability. For Cases 1 and 2, four sub-scenarios were developed to represent different expert opinion conditions: (a) the physician is always correct, (b) the physician is always incorrect, (c) the physician consistently answers 'yes', and (d) the physician consistently answers 'no'. In Case 3, we simulate physicians with different expertise by varying the probability p of providing a correct answer, with $p \in \{70\%, 75\%, 80\%, 85\%, 90\%, 95\%\}$. The LLM's role was to assess the physician's input and provide its own binary response, accompanied by an explanation. UNITN also explored

	Scenario 1			Scenario 2			Scenario 3			Scenario 4		
	Med	Ll2	Mis	Med	Ll2	Mis	Med	Ll2	Mis	Med	Ll2	Mis
1a	22	**84**	43	**97**	96	70	54	**91**	47	**85**	83	66
1b	79	57	**95**	7	14	**85**	51	19	**95**	37	52	**90**
1c	38	70	**75**	66	71	**80**	49	70	**77**	62	70	**82**
1d	62	**71**	64	38	40	**74**	55	40	**65**	60	65	**74**

Table 8: Accuracy (in %) in Case 1.

	Scenario 1			Scenario 2			Scenario 3			Scenario 4		
	Med	Ll2	Mis	Med	Ll2	Mis	Med	Ll2	Mis	Med	Ll2	Mis
1a	**32**	7	9	**23**	6	7	23	**26**	7	23	**26**	9
1b	**32**	28	30	**20**	8	8	**23**	10	8	23	21	**28**
1c	**32**	7	9	**23**	6	16	23	**26**	16	23	**26**	25
1d	**32**	28	9	20	**28**	8	**23**	10	8	23	21	**28**

Table 9: ROUGE-L scores in Case 1.

whether the order of few-shot examples (simulated interactions between the physician and the LLM) impacts the LLM's performance; Scenarios 1–4 are organized to introduce variability in the sequence of agreement or disagreement on 'yes' or 'no' responses between the physician and the LLM. The left side of Figure 12 shows the prompt template; it starts with task instructions, followed by few-shot examples, leading to the test question, which includes the relevant context and the expert's opinion. The LLM's response was triggered by the keyword "Assistant:".

The experimental study addressed the following research questions:

- **Q1**: Can LLMs correct physicians when needed?
- **Q2**: Can LLMs explain the reasons behind their answers?
- **Q3**: Can LLMs correct physicians when they provide arguments for their answers?
- **Q4**: Can LLMs fed with physician answers outperform both themselves and physicians?

For this investigation, a binary subset comprising 445 test examples—62% labeled as "yes"—was extracted from the PubMedQA dataset [Jin et al., 2019]. GPT-4 [Team, 2024] was then employed to generate both plausible correct and incorrect explanations for each test instance, simulating the explanations physicians might offer in Case 2. UNITN has made this enriched dataset available online[1]. Three models were utilized in the experiments: Meditron-7B (Med) [Chen et al., 2023a, Chen et al., 2023b], Llama2-7B chat (Ll2) [Touvron et al., 2023], and Mistral-7B-Instruct (Mis) [Jiang et al., 2023]. The source code can be accessed online[2].

	Scenario 1			Scenario 2			Scenario 3			Scenario 4		
	Med	Ll2	Mis	Med	Ll2	Mis	Med	Ll2	Mis	Med	Ll2	Mis
2a	46	**97**	96	89	**99**	95	2	**99**	87	**100**	98	95
2b	**99**	0	56	30	0	**50**	**93**	2	60	4	5	**51**
2c	39	61	**72**	**79**	62	69	40	63	**69**	63	62	**69**
2d	62	38	**78**	37	37	**76**	63	39	**86**	39	41	**79**

Table 10: Accuracy (in %) in Case 2

	Scenario 1			Scenario 2			Scenario 3			Scenario 4		
	Med	Ll2	Mis	Med	Ll2	Mis	Med	Ll2	Mis	Med	Ll2	Mis
2a	14	28	**31**	**37**	36	21	37	**38**	34	26	**39**	30
2b	14	**24**	19	13	13	**22**	24	15	**26**	13	15	**40**
2c	14	11	**25**	**37**	11	22	37	**38**	30	26	**34**	30
2d	14	**24**	19	13	13	**22**	24	15	**26**	13	15	**40**

Table 11: ROUGE-L scores in Case 2

[1] https://tinyurl.com/pubmedqa-with-gpt4-exp
[2] https://tinyurl.com/physician-medLLM-interaction

4.2.3 Results

1. Prompt design affects LLM performance in correcting erroneous physician responses. Table 8 shows the remarkable influence of prompt design on the models' performances; given appropriate instructions and examples, LLMs can effectively correct physicians. For instance, in Case 1d, the physician always responds with "no" while the ground truth distribution of class "no" is just 38%: Mistral achieves significantly higher accuracy, while Llama2 and Meditron exhibit greater sensitivity to prompt changes, displaying improved performance in Scenarios 1 and 4.

2. LLMs *could* explain reasons behind their answers. Detailed responses from each model in Case 1 showed that the quality of Meditron's explanations exhibits minimal sensitivity to the physician's short answer (see Table 9). Llama2 model typically yields lower ROUGE-L scores in cases 1a (the physician is always right) and 1c (the physician always says "yes"). Conversely, the Mistral model consistently delivers better explanations in Scenario 4 for cases b, c, and d. Overall, results show that LLMs are capable of generating plausible explanations when the prompt is constructively framed.

	1b-4	**2b-4**
Ground Truth	An association was found between sublingual varices and hypertension. Examining the lateral borders of the tongue is easily done, causes no harm and could be a valuable method for the dental profession to take active part in preventive healthcare.	An association was found between sublingual varices and hypertension. Examining the lateral borders of the tongue is easily done, causes no harm and could be a valuable method for the dental profession to take active part in preventive healthcare.
Med	Sublingual varices are not related to hypertension.	The study found a correlation between sublingual varices and hypertension, but it is possible to argue that this correlation does not necessarily imply causation. For instance, it could be hypothesized that both conditions are independently associated with another common factor such as age or lifestyle habits, rather than sublingual varices being directly linked to hypertension. This alternative interpretation, while not supported by the data, suggests that the presence of sublingual varices does not necessarily correspond to hypertension.
Ll2	Sublingual varices are associated with hypertension, with a positive predictive value of 0.5 and a negative predictive value of 0.80.	The study found a correlation between sublingual varices and hypertension, but it is possible that this correlation is due to a third factor rather than a direct causal relationship between the two conditions.
Mis	Sublingual varices are a reliable indicator of hypertension, with a positive predictive value of 0.5 and a negative predictive value of 0.80.	Sublingual varices and hypertension are correlated, but it is unclear whether this association is causal. Other factors, such as age or lifestyle habits, may be responsible for both conditions.

Table 12: Example of explanations generated by each model on different cases.

3. LLMs exhibit different levels of dependence on physician-provided arguments.

	Scenario 1			Scenario 2			Scenario 3			Scenario 4		
	Med	Ll2	Mis	Med	Ll2	Mis	Med	Ll2	Mis	Med	Ll2	Mis
Baseline	81	80	**84**	83	81	84	85	79	84	**84**	79	**84**
Phy_70	40	75	58	70	71	74	55	69	61	71	75	73
Phy_75	35	77	58	74	75	74	54	73	61	71	74	72
Phy_80	34	80	55	79	80	74	51	76	56	79	78	72
Phy_85	28	80	52	85	84	72	53	80	55	80	78	70
Phy_90	28	80	49	88	87	71	54	83	52	79	80	69
Phy_95	24	82	46	**92**	**92**	71	53	**87**	49	82	81	67

Table 13: Accuracy of models in Case 3

Table 10 shows LLMs tend to depend heavily on physician-provided arguments, and LLMs are influenced by the order and content of few-shot examples. Meditron, for instance, reaches 100% accuracy in Case 2a, Scenario 4, where the physician provides correct answers with accurate explanations, suggesting a preference to focus on the latest examples in the prompt. This is reflected in its strong performance in Scenarios 2 and 4. Conversely, in Case 2b, where the physician's answers oppose the ground truth, Meditron performs better in Scenarios 1 and 3, learning to contradict the physician responses. Meanwhile, LLama2 often relies on the physician's input across all scenarios. In contrast, Mistral shows robust performance, minimally affected by prompt changes and maintaining over 75% accuracy in Case 2d, demonstrating its capability to effectively correct incorrect physician answers. Table 11 shows the ROUGE-L scores in Case 2; LLama2 and Mistral provide more plausible and detailed explanations when the prompt has the physician's opinion. Meditron heavily relies on the physician's input, adversely affecting the quality of its explanations. Table 12 shows a ground truth explanation and the model explanations. Meditron adapts to the physician's input, while Llama2 and Mistral remain consistent, offering reasonable explanations regardless of input.

4. LLMs improve with expert answers but fail to outperform them. Table 13 the Case 3 results; the baseline performance of models is consistent across scenarios. Meditron improves in Scenarios 2 and 4, but only surpasses its baseline in Scenario 2 when physician accuracy exceeds 80%. LLama2 exceeds its baseline in all scenarios with physician accuracies over 85%. Conversely, Mistral performs poorly, significantly influenced by the physician's answers. All in all, models improve when interacting with physicians but fail to outperform them.

Overall, this study provides significant insights into the application of LLMs in medical decision support; UNITN demonstrated that the effectiveness of LLMs in correcting physician errors is highly dependent on the clarity and specificity of instructions and examples provided in the prompt. Moreover, LLMs' capacity to elucidate their reasoning is also contingent upon these well-crafted prompts. The research highlights that while LLMs generally rely on physicians' arguments, their performance can be skewed by the order of few-shot examples introduced in the prompts. Acknowledging its limitations, the study revealed that its reliance on GPT-4 to simulate plausible and incorrect responses necessitates real-world validation; engaging actual physician interactions is essential for further authenticating and refining the findings. The primary objective of this research was to underscore the limitations of current open-source LLMs in medical support rather than to propose conclusive solutions. By highlighting

these critical gaps, UNITN aims to stimulate further research focused on overcoming these challenges, ultimately advancing the role of LLMs in enhancing medical decision-making.

5 Ethical Considerations

For the ethical and legal risks that arise from the use of AI in decision-making processes, we refer to the ethical considerations presented in D2.1. This section outlines the challenges associated with the TANGO systems, emphasizing the need to assess and mitigate risks related to unfair behavior of explainable AI models, privacy violations, and unacceptable human persuasion.

Fairness AI models often inherit biases from training data sourced from open and uncurated datasets, leading to discriminatory or harmful outputs. Even explanation models may perpetuate such biases and some studies observed that unfortunately fair models do not necessarily produce *fair explanations* [Yang and Howe, 2024]. As a consequence, it is fundamental to design and develop methods for deriving fair explanation processes [Zhao et al., 2023]. In other words, explanations should provide insights that do not disadvantage some social groups against others perpetuating or amplifying existing social inequalities. Particular attention should be posed on the risk of reinforcing systemic biases and disparities against well-known underrepresented groups, such as women and marginalized genders, people from a low income background and/or people of different ethnic backgrounds, such as migrants. We remark that when providing suggestions for actionable recourse – guiding individuals on how to change a model's prediction – it is critical to ensure that these explanations are designed to require an effort that is both reasonable and equitable, considering the diverse capabilities, resources, and opportunities available to different individuals and groups under analysis. This means that recommendations must account for variations in socioeconomic status, educational background, access to resources, and systemic barriers that may disproportionately affect certain individuals and populations. The goal is to prevent scenarios where individuals from less advantaged groups face disproportionate challenges in implementing the suggested actions compared to those with greater access to opportunities, ensuring that the pathway to achieving favorable outcomes remains fair and accessible to all. In such scenarios it is also important to consider the issue of *intersectional fairness*, i.e., forms of discrimination towards groups of people with different, sensitive characteristics (e.g. gender, race, age, sexuality, religion, disability, etc.). In TANGO we have already some results in terms of fairness guarantees in hybrid decision making, thanks to the development of an interpretable fair abstention classifier [Lenders et al., 2024]. Despite the efforts of the scientific community on the field of fairness in AI, ensuring fairness in both decisions and explanations remains a challenge. Particularly so when using large language models (LLMs) to generate natural language explanations, which may contain biased or inappropriate language targeting underrepresented groups.

Privacy Privacy concerns arise from the use of non-anonymized data, lack of privacy-preserving learning strategies, and the transparency provided by explainers, which can inadvertently reveal sensitive information. TANGO addresses these concerns by developing a framework to assess privacy exposure in AI models and explainers [Naretto et al., 2024] and implementing mitigation techniques, such as differential privacy, potentially combined with computation processes such as federated learning, that can facilitate privacy protection by avoiding data sharing [Haffar et al., 2024].

Persuasion LLMs, trained on vast textual datasets, possess the ability to generate coherent and persuasive content. This raises concerns about their potential to manipulate opinions, spread misinformation, and exacerbate social issues like political polarization. In the TANGO context, persuasive AI explanations could convince a human to make decisions that do not align with their own preferences and goals. Appropriate awareness strategies should be put in place to allow humans to prevent these undesired outcomes.

6 Conclusions

This deliverable has discussed the TANGO framework for the design of Interactive Decision-making process, exploiting the cognitive theory introduced in WP1 and presented in D1.1. The framework leverages the Situation Aware model as a core component to enhance human-AI collaboration, focusing on improving explainability and fostering trust, aiming to create a deeper alignment between human operators and AI, emphasizing mutual understanding at every stage of interaction.

The document has presented the results of the research activities within WP3, focusing on the design and evaluation of explanation processes involving both visual and text-based explanation. In particular, visual-based explanations exploits the encoding of information in the form of graphs or image representation, to provide the user a direct channel to understand the AI decision-making process. Text-based explanations, on the other hand, leverage the natural language processing capabilities of AI to provide a more detailed and human-like explanation of the decision-making process.

References

[Alkaissi and McFarlane, 2023] Alkaissi, H. and McFarlane, S. I. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*, 15(2):e35179.

[Azamfirei et al., 2023] Azamfirei, R., Kudchadkar, S. R., and Fackler, J. (2023). Large language models and the perils of their hallucinations. *Critical Care*, 27(1):1–2.

[Bronzini et al., 2024] Bronzini, M., Nicolini, C., Lepri, B., Staiano, J., and Passerini, A. (2024). Unveiling llms: The evolution of latent representations in a dynamic knowledge graph. In *First Conference on Language Modeling*. OpenReview.net.

[Cappuccio et al., 2024] Cappuccio, E., Fadda, D., Lanzilotti, R., and Rinzivillo, S. (2024). Fiper: a visual-based explanation combining rules and feature importance. *CoRR*, abs/2404.16903.

[Chen et al., 2023a] Chen, Z., Hernández-Cano, A., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., Pagliardini, M., Fan, S., Köpf, A., Mohtashami, A., Sallinen, A., Sakhaeirad, A., Swamy, V., Krawczuk, I., Bayazit, D., Marmet, A., Montariol, S., Hartley, M.-A., Jaggi, M., and Bosselut, A. (2023a). Meditron-70b: Scaling medical pretraining for large language models. *arXiv*, abs/2311.16079.

[Chen et al., 2023b] Chen, Z., Hernández-Cano, A., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., Pagliardini, M., Fan, S., Köpf, A., Mohtashami, A., Sallinen, A., Sakhaeirad, A., Swamy, V., Krawczuk, I., Bayazit, D., Marmet, A., Montariol, S., Hartley, M.-A., Jaggi, M., and Bosselut, A. (2023b). Meditron-70b: Scaling medical pretraining for large language models.

[Chromik and Butz, 2021] Chromik, M. and Butz, A. (2021). Human-xai interaction: A review and design principles for explanation user interfaces. In Ardito, C., Lanzilotti, R., Malizia, A., Petrie, H., Piccinno, A., Desolda, G., and Inkpen, K., editors, *Human-Computer Interaction - INTERACT 2021 - 18th IFIP TC 13 International Conference, Bari, Italy, August 30 -*

September 3, 2021, Proceedings, Part II, volume 12933 of *Lecture Notes in Computer Science*, pages 619–640. Springer.

[Dhurandhar et al., 2018] Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., and Das, P. (2018). Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *Advances in neural information processing systems*, 31.

[Dhurandhar et al., 2019] Dhurandhar, A., Pedapati, T., Balakrishnan, A., Chen, P.-Y., Shanmugam, K., and Puri, R. (2019). Model agnostic contrastive explanations for structured data. *arXiv preprint arXiv:1906.00117*.

[Diggelmann et al., 2020] Diggelmann, T., Boyd-Graber, J. L., Bulian, J., Ciaramita, M., and Leippold, M. (2020). CLIMATE-FEVER: A dataset for verification of real-world climate claims. *CoRR*, abs/2012.00614.

[Endsley, 1995] Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Hum. Factors*, 37(1):32–64.

[Farid et al., 2023] Farid, K., Schrodi, S., Argus, M., and Brox, T. (2023). Latent diffusion counterfactual explanations. *arXiv preprint arXiv:2310.06668*.

[Fernández-Loría et al., 2021] Fernández-Loría, C., Provost, F., and Han, X. (2021). Explaining data-driven decisions made by ai systems: The counterfactual approach.

[Ghandeharioun et al., 2024] Ghandeharioun, A., Caciularu, A., Pearce, A., Dixon, L., and Geva, M. (2024). Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

[Guidotti, 2022] Guidotti, R. (2022). Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, pages 1–55.

[Guidotti et al., 2019a] Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., and Turini, F. (2019a). Factual and counterfactual explanations for black box decision making. *IEEE Intell. Syst.*, 34(6):14–23.

[Guidotti et al., 2019b] Guidotti, R., Monreale, A., Matwin, S., and Pedreschi, D. (2019b). Black box explanation by learning image exemplars in the latent feature space. In Brefeld, U., Fromont, É., Hotho, A., Knobbe, A. J., Maathuis, M. H., and Robardet, C., editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2019, Würzburg, Germany, September 16-20, 2019, Proceedings, Part I*, volume 11906 of *Lecture Notes in Computer Science*, pages 189–205. Springer.

[Gunning and Aha, 2019] Gunning, D. and Aha, D. W. (2019). Darpa's explainable artificial intelligence (XAI) program. *AI Mag.*, 40(2):44–58.

[Haffar et al., 2024] Haffar, R., Naretto, F., Sánchez, D., Monreale, A., and Domingo-Ferrer, J. (2024). GLOR-FLEX: local to global rule-based explanations for federated learning. In *FUZZ*, pages 1–9. IEEE.

[Harary and Gupta, 1997] Harary, F. and Gupta, G. (1997). Dynamic graph models. *Mathematical and Computer Modelling*, 25(7):79–87.

[Hwang et al., 2022] Hwang, J., Lee, T., Lee, H., and Byun, S. (2022). A clinical decision support system for sleep staging tasks with explanations from artificial intelligence: User-centered design and evaluation study. *J Med Internet Res*, 24(1):e28659.

[Ji et al., 2023] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):248:1–248:38.

[Jiang et al., 2023] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7b. *arXiv*, abs/2310.06825.

[Jin et al., 2019] Jin, Q., Dhingra, B., Liu, Z., Cohen, W., and Lu, X. (2019). Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

[Kanamori et al., 2020] Kanamori, K., Takagi, T., Kobayashi, K., and Arimura, H. (2020). Dace: Distribution-aware counterfactual explanation by mixed-integer linear optimization. In *IJCAI*, pages 2855–2862.

[Lenders et al., 2024] Lenders, D., Pugnana, A., Pellungrini, R., Calders, T., Pedreschi, D., and Giannotti, F. (2024). Interpretable and fair mechanisms for abstaining classifiers. In *ECML/PKDD (7)*, volume 14947 of *Lecture Notes in Computer Science*, pages 416–433. Springer.

[Makhzani et al., 2015] Makhzani, A., Shlens, J., Jaitly, N., and Goodfellow, I. J. (2015). Adversarial autoencoders. *CoRR*, abs/1511.05644.

[Miller, 2019] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38.

[Molnar, 2022] Molnar, C. (2022). *Interpretable Machine Learning*. Independently published, 2 edition.

[Naretto et al., 2024] Naretto, F., Monreale, A., and Giannotti, F. (2024). Evaluating the privacy exposure of interpretable global and local explainers. *Accepted at Trans. Data Priv.*

[Poyiadzi et al., 2020] Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., and Flach, P. (2020). Face: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350.

[Salvagno et al., 2023] Salvagno, M., Taccone, F., Gerli, A., and ChatGPT (2023). Can artificial intelligence help for scientific writing? *Critical Care*, 27(1):75. Publisher: BioMed Central Ltd.

[Sayin et al., 2024] Sayin, B., Minervini, P., Staiano, J., and Passerini, A. (2024). Can LLMs correct physicians, yet? investigating effective interaction methods in the medical domain. In Naumann, T., Ben Abacha, A., Bethard, S., Roberts, K., and Bitterman, D., editors, *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 218–237, Mexico City, Mexico. Association for Computational Linguistics.

[Schneider, 2024] Schneider, J. (2024). Explainable generative ai (genxai): A survey, conceptualization, and research agenda. *arXiv preprint arXiv:2404.09554*.

[Springer and Whittaker, 2020] Springer, A. and Whittaker, S. (2020). Progressive disclosure: When, why, and how do users want algorithmic transparency information? *ACM Trans. Interact. Intell. Syst.*, 10(4):29:1–29:32.

[Sutton et al., 2020] Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., and Kroeker, K. I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digital Medicine*, 3(1):17.

[Team, 2024] Team, O. (2024). Gpt-4 technical report. *arXiv*, abs/2303.08774.

[Thorne et al., 2018] Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). FEVER: a large-scale dataset for fact extraction and verification. In Walker, M. A., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 809–819. Association for Computational Linguistics.

[Touvron et al., 2023] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv*, abs/2307.09288.

[Wachter et al., 2017] Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841.

[Wan et al., 2018] Wan, W., Zhong, Y., Li, T., and Chen, J. (2018). Rethinking feature distribution for loss functions in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9117–9126.

[Wang et al., 2017] Wang, Q., Mao, Z., Wang, B., and Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE transactions on knowledge and data engineering*, 29(12):2724–2743.

[Yang et al., 2023] Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., and Ni, B. (2023). Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41.

[Yang and Howe, 2024] Yang, Y. and Howe, B. (2024). Does a fair model produce fair explanations? relating distributive and procedural fairness. In *HICSS*, pages 6868–6877. ScholarSpace.

[Zhang and Nanda, 2024] Zhang, F. and Nanda, N. (2024). Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

[Zhao et al., 2023] Zhao, Y., Wang, Y., and Derr, T. (2023). Fairness and explainability: Bridging the gap towards fair model explanations. In *AAAI*, pages 11363–11371. AAAI Press.

[Zheng and Sun, 2019] Zheng, Z. and Sun, L. (2019). Disentangling latent space for vae by label relevant/irrelevant dimensions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12192–12201.