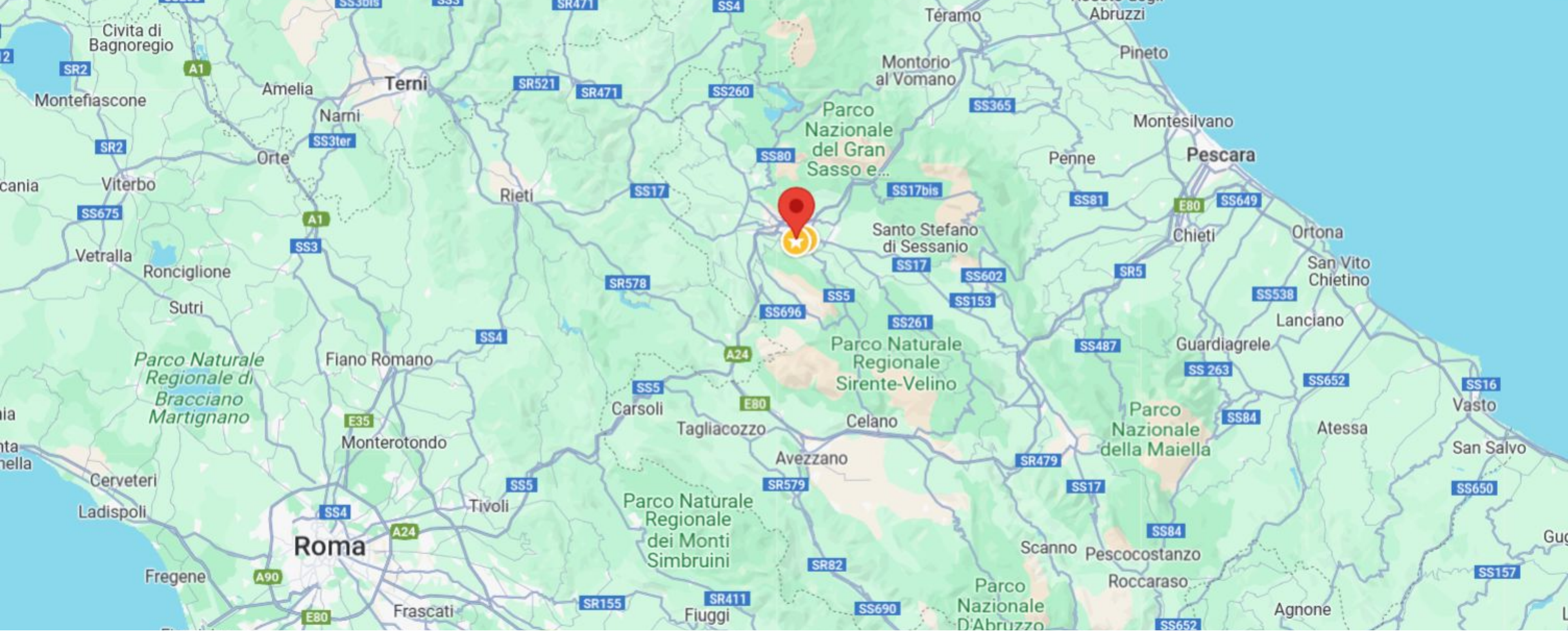
Exploring Large Language Models Capabilities to Explain Decision Trees

Paulo Bruno De Sousa Serafim

08/Feb/2024

GRAN SASSO SCIENCE INSTITUTE
SCHOOL OF ADVANCED STUDIES
Scuola Universitaria Superiore

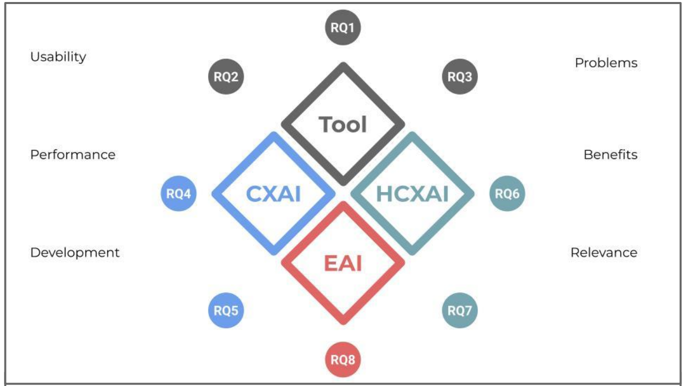


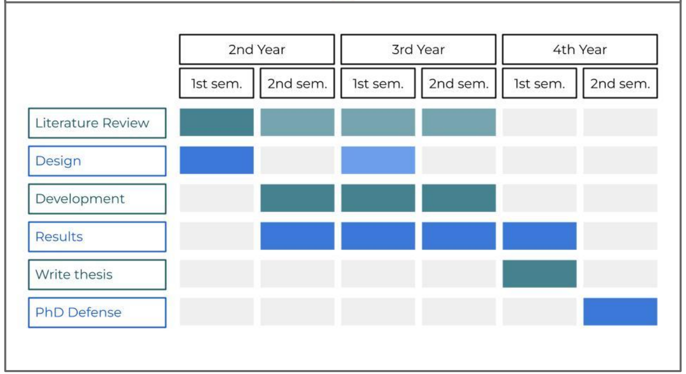
Supervisors:

Pierluigi Crescenzi - *Gran Sasso Science Institute*

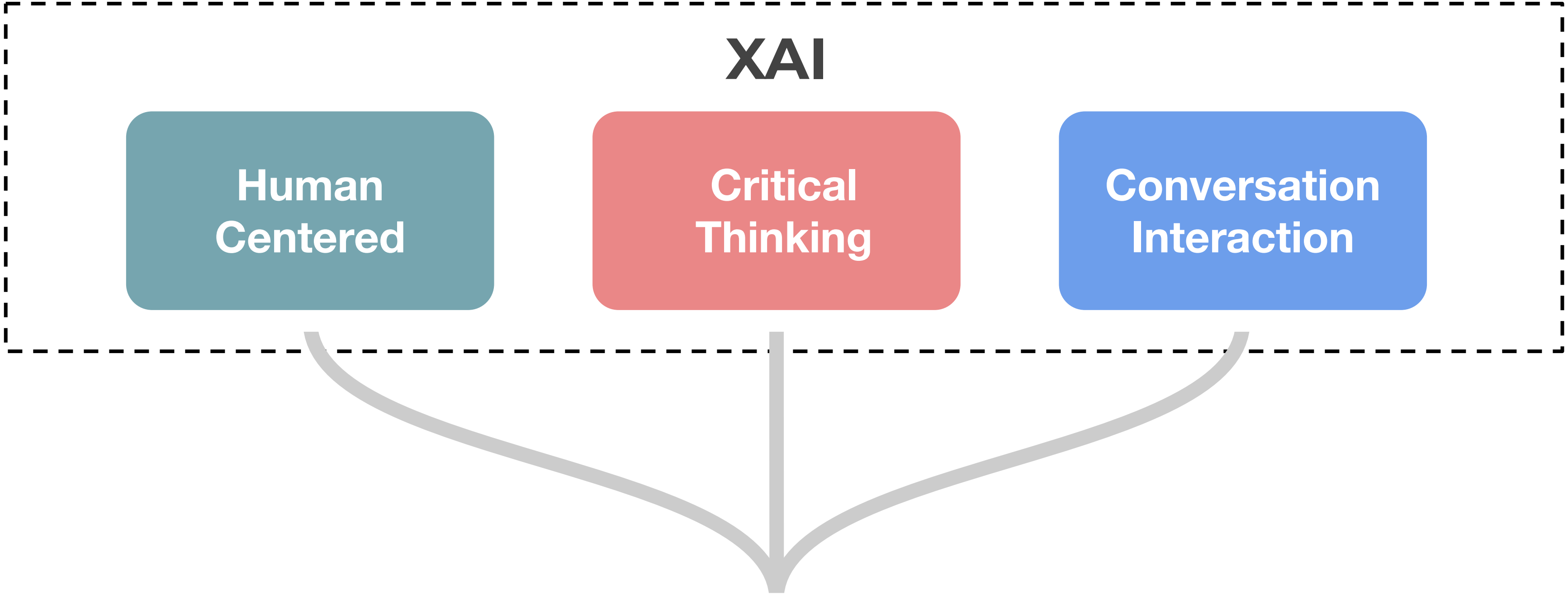
Fosca Giannotti - *Scuola Normale Superiore di Pisa*

“Development of a User-Centered Conversational Explainable AI Tool”





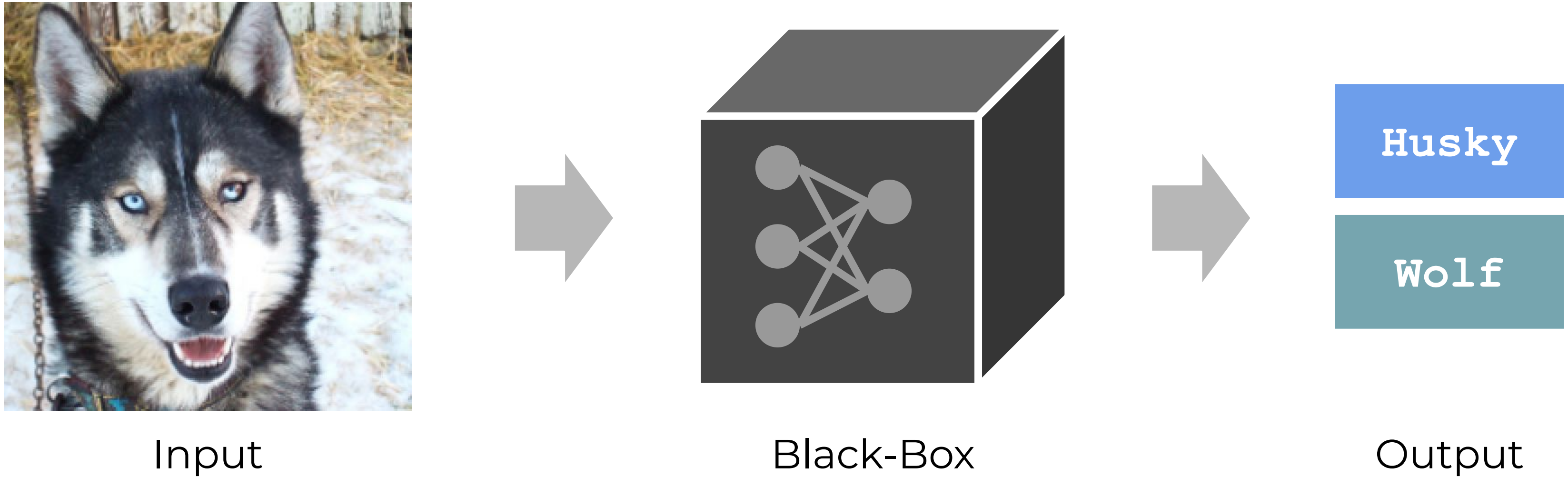
Thesis proposal presentation - 6 Oct 2023



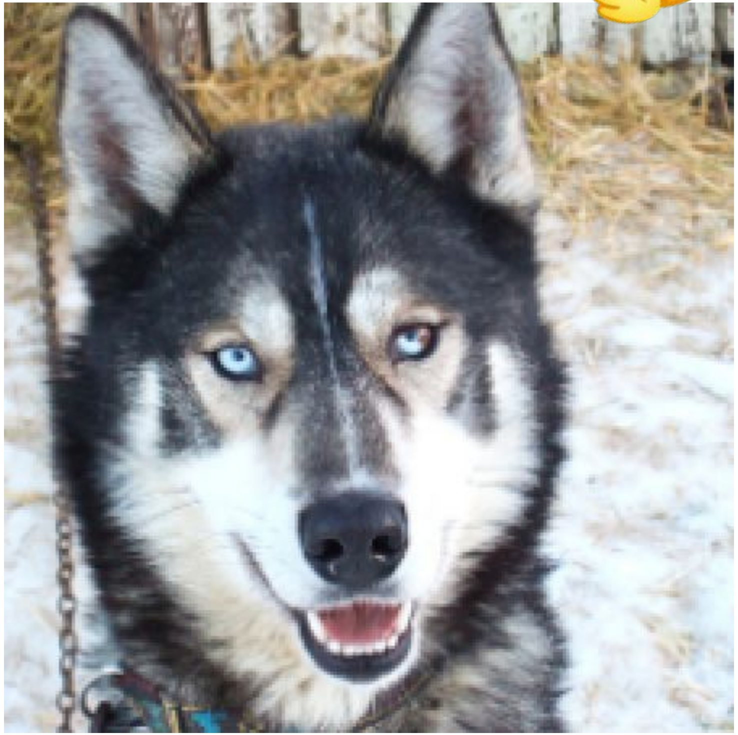
“Development of a User-Centered Conversational Explainable AI Tool”

Explainable
Artificial Intelligence

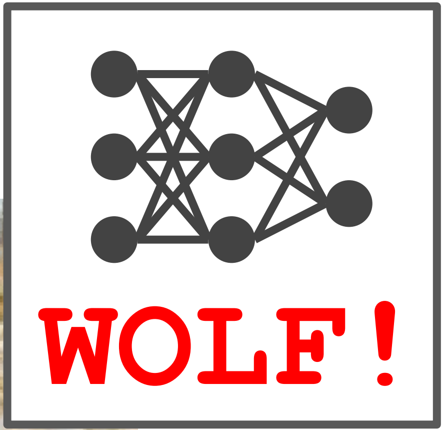
“XAI”


Husky
Wolf
Input
Black-Box
Output

Doggy



~~Doggy~~



Wolf?

Wolf

NTSB National Transportation Safety Board

Investigations Safety Research News & Events Advocacy Family Assistance About Us

Home > Investigations > Collision Between Vehicle Cont...

Completed Investigatio

Investigation No
HWY18MH010

Event Date
3/18/2018

Location
Tempe, AZ

USA

Family Assistance Contact
assistance@ntsb.gov

NTSB Media Relations
ntsbmediarelations@ntsb.gov

Docket
- HWY18MH010

REUTERS World Business Markets Breakingviews

RETAIL OCTOBER 11, 2018 / 1:04 AM / UPDATED 5 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like wo

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

RESEARCH

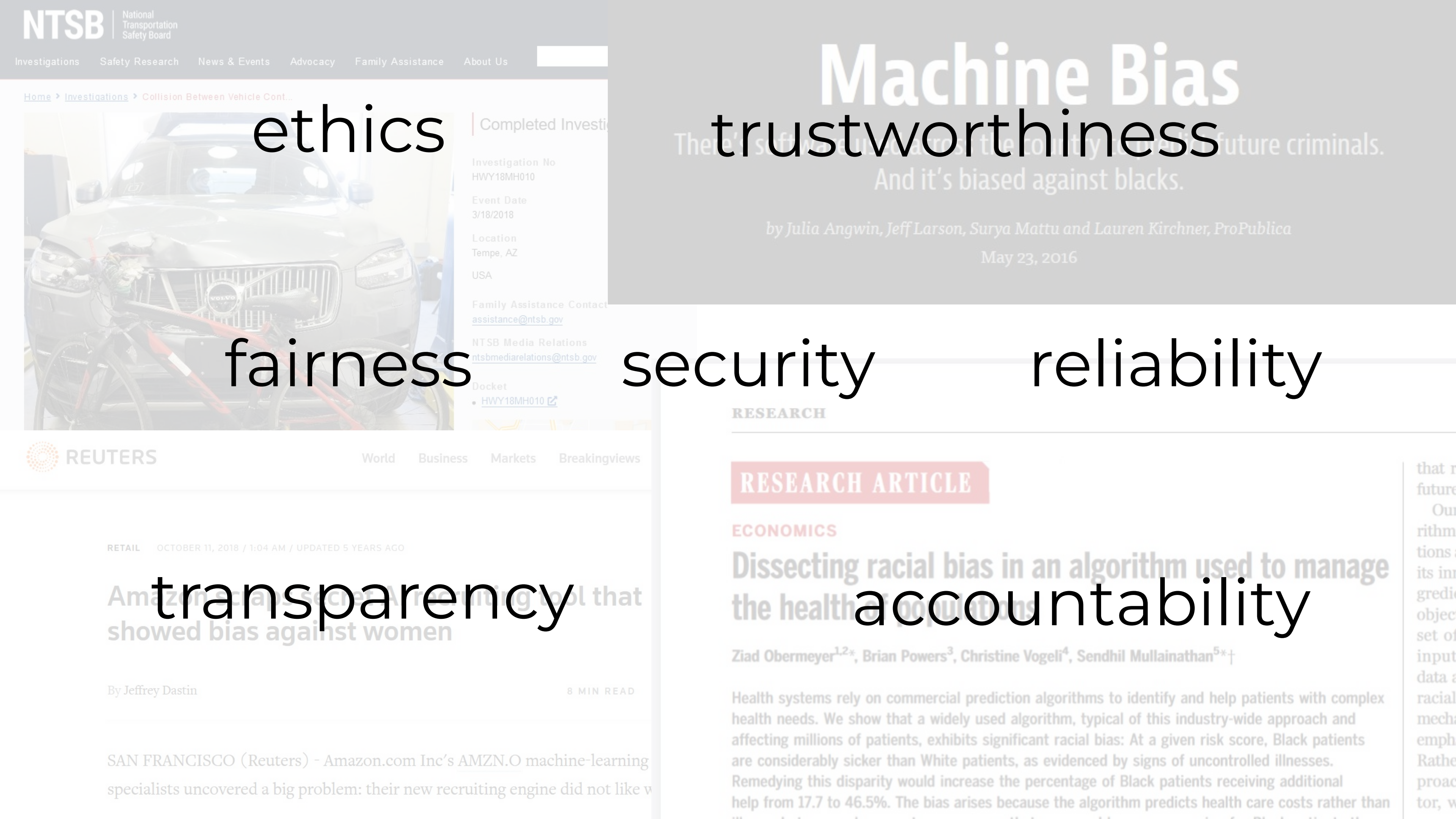
RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2*], Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5*†]

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than



ethics
trustworthiness
fairness
security
reliability
transparency
accountability
NTSB National Transportation Safety Board
Investigations Safety Research News & Events Advocacy Family Assistance About Us
Machine Bias
And it's biased against blacks.
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016
REUTERS
showed bias against women
By Jeffrey Dastin
RESEARCH
RESEARCH ARTICLE
ECONOMICS
Dissecting racial bias in an algorithm used to manage
Ziad Obermeyer, Brian Powers, Christine Vogeli, Sendhil Mullainathan

How to use XAI methods?

- Programming

- [ ] **Programming**
- [ ] **Machine Learning**

- ▢ Programming
- ▢ Machine Learning
- ▢ Python

- [ ] Programming
- [ ] Machine Learning
- [ ] Python
- [ ] PyTorch/TensorFlow

⋮

- [ ] Programming
- [ ] Machine Learning
- [ ] Python
- [ ] PyTorch/TensorFlow

⋮

ML Expert

“Make XAI human-centered”

Interactive
XAI!

"Make XAI human-centered"

What can we use to improve

user interaction?

What can we use to improve

user interaction?

ChatGPT

Conversational AI

Bard

Gemini

Conversational AI

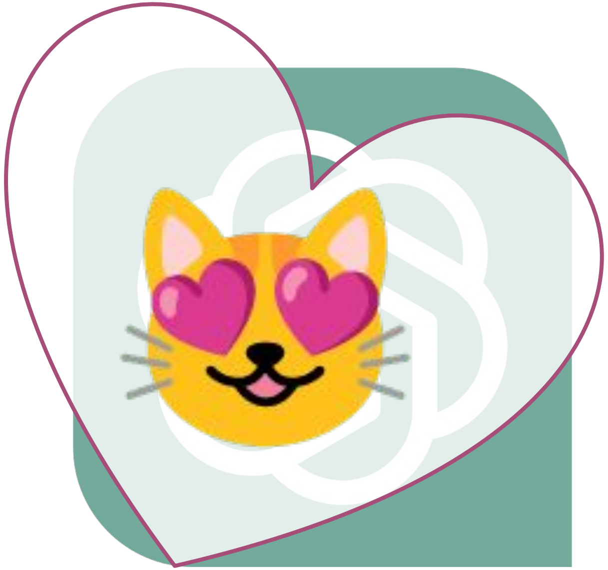
Why?

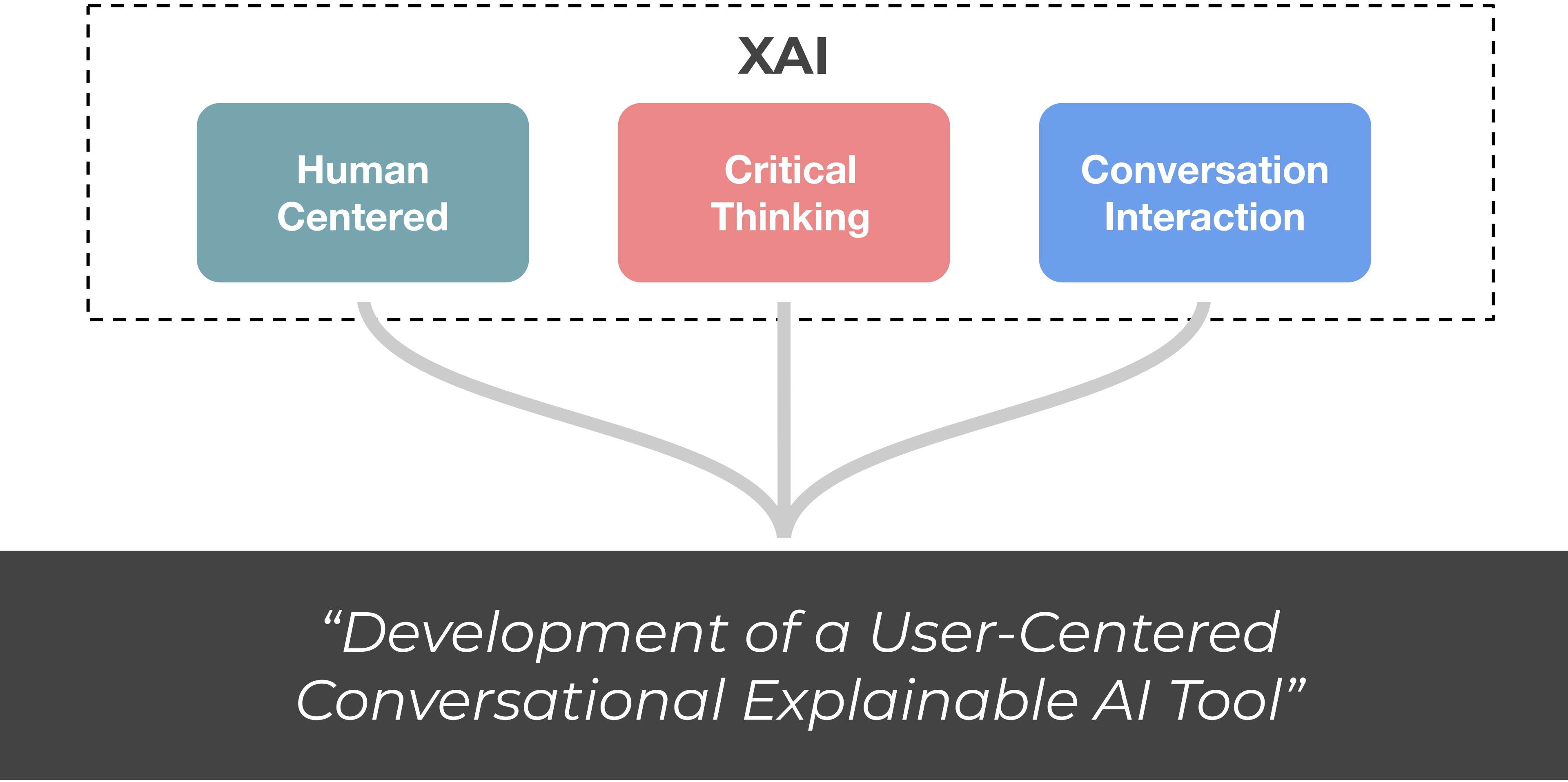
Stanford Alpaca

Llama

PaLM 2

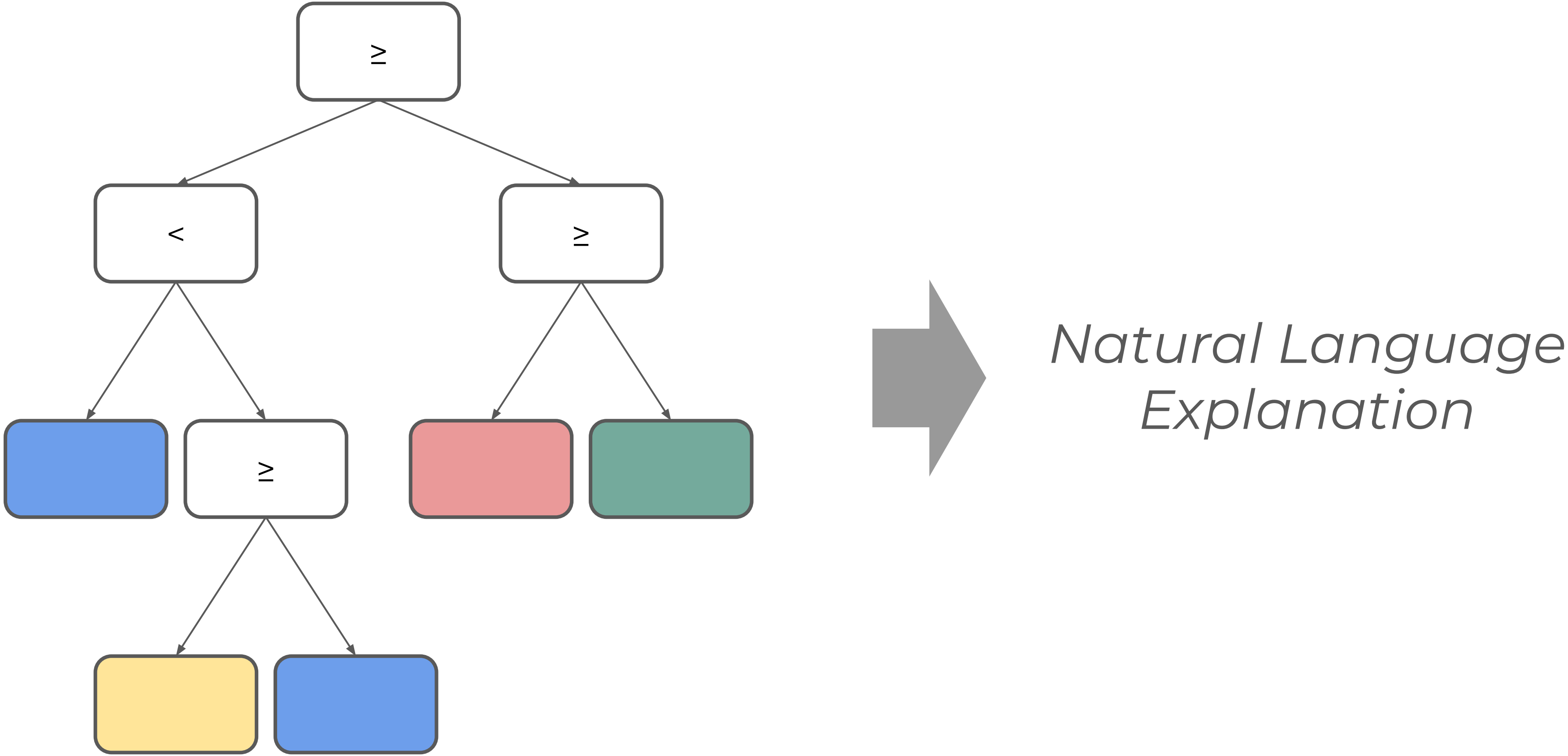
python™

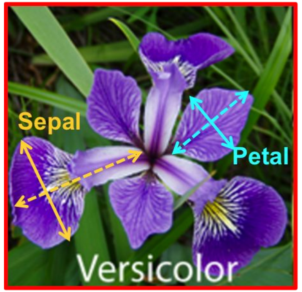



XAI
Human Centered
Critical Thinking
Conversation Interaction
“Development of a User-Centered Conversational Explainable AI Tool”

Exploring Large Language Models Capabilities to Explain Decision Trees

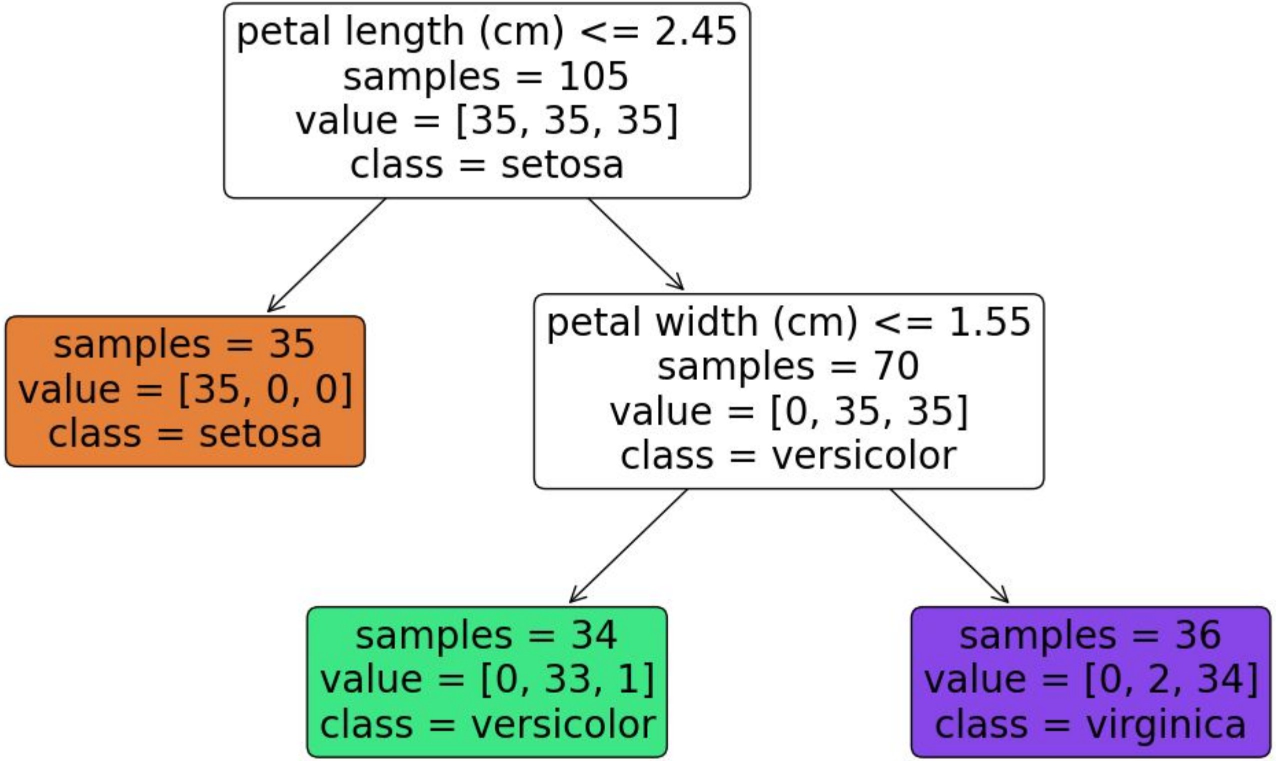
Feb/2024

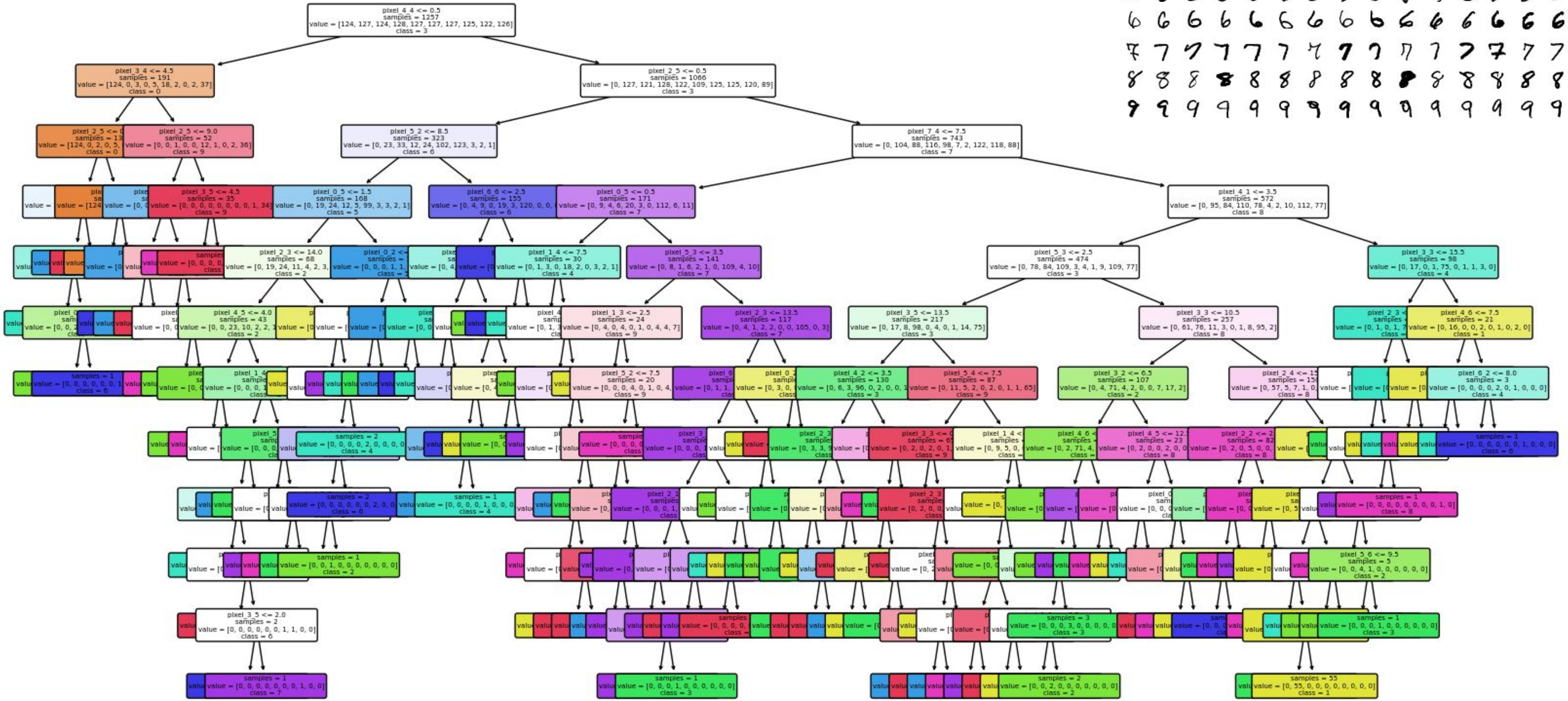

≥
<
≥
≥
Natural Language
Explanation

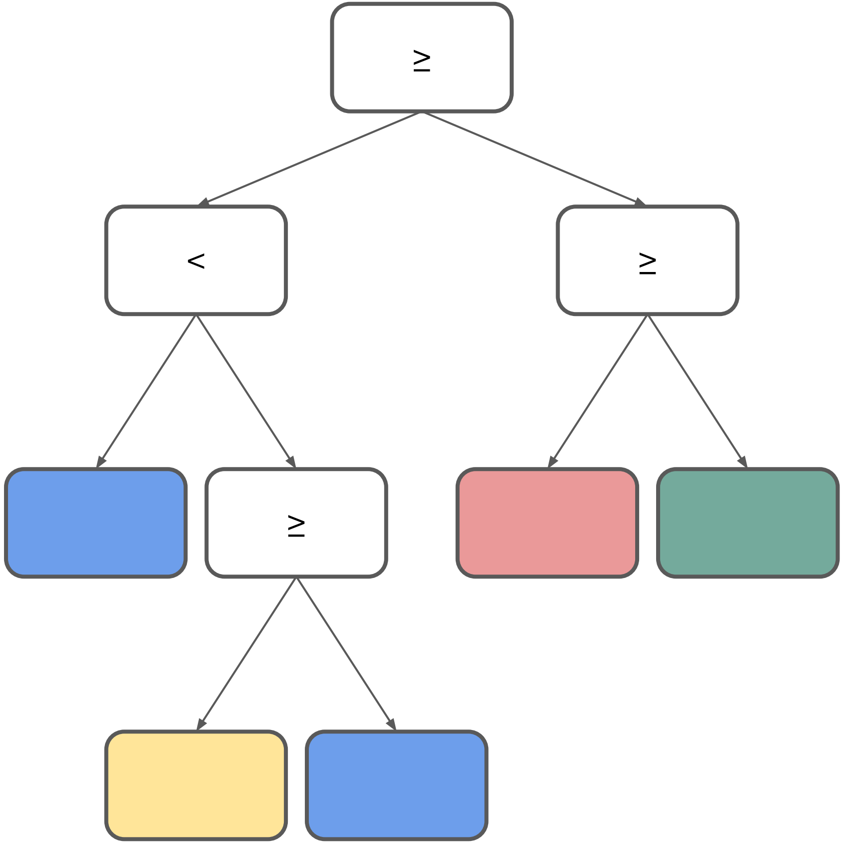

Sepal
Petal
Versicolor


Virginica


Setosa


petal length (cm) <= 2.45
samples = 105
value = [35, 35, 35]
class = setosa
samples = 35
value = [35, 0, 0]
class = setosa
petal width (cm) <= 1.55
samples = 70
value = [0, 35, 35]
class = versicolor
samples = 34
value = [0, 33, 1]
class = versicolor
samples = 36
value = [0, 2, 34]
class = virginica


pixel_4_4 <= 0.5
samples = 1257
value = [124, 127, 124, 128, 127, 127, 127, 125, 122, 126]
class = 3
pixel_3_4 <= 4.5
samples = 191
value = [124, 0, 3, 0, 5, 18, 2, 0, 2, 37]
class = 0
pixel_2_5 <= 0.5
samples = 1066
value = [0, 127, 121, 128, 122, 109, 125, 125, 120, 89]
class = 3
pixel_5_2 <= 8.5
samples = 323
value = [0, 23, 33, 12, 24, 102, 123, 3, 2, 1]
class = 6
pixel_7_4 <= 7.5
samples = 743
value = [0, 104, 88, 116, 98, 7, 2, 122, 118, 88]
class = 7
pixel_2_5 <= 9.0
samples = 52
value = [0, 0, 1, 0, 0, 12, 1, 0, 2, 36]
class = 9
pixel_0_5 <= 1.5
samples = 168
value = [0, 19, 24, 12, 5, 99, 3, 3, 2, 1]
class = 5
pixel_0_5 <= 0.5
samples = 171
value = [0, 9, 4, 6, 20, 3, 0, 112, 6, 11]
class = 7
pixel_4_1 <= 3.5
samples = 572
value = [0, 95, 84, 110, 78, 4, 2, 10, 112, 77]
class = 8
pixel_5_3 <= 2.5
samples = 474
value = [0, 78, 84, 109, 3, 4, 1, 9, 109, 77]
class = 3
pixel_3_3 <= 15.5
samples = 98
value = [0, 17, 0, 1, 75, 0, 1, 1, 3, 0]
class = 4



Natural Language Explanation

Generación Automática de Explicaciones en Lenguaje Natural para Árboles de Decisión de Clasificación

B. López-Trigo, Jose M. Alonso, A. Bugarín
Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS),
Universidade de Santiago de Compostela,
Campus Vida, E-15782, Santiago de Compostela, Spain
Email: bruno.lopez.trigo@rai.usc.es, {josemaria.alonso.moral, alberto.bugarin.diz}@usc.es

***Resumen*—En este trabajo describimos un modelo de explicaciones en lenguaje natural para árboles de decisión para clasificación. Las explicaciones incluyen aspectos globales del clasificador y aspectos locales de la clasificación de una instancia concreta. La propuesta está implementada en el servicio Web de código abierto ExpliClas [1], que en su versión actual opera sobre árboles construidos con Weka y conjuntos de datos con atributos numéricos. Ilustramos la viabilidad de la propuesta con dos casos de ejemplo, donde mostramos paso a paso cómo el modelo explica los respectivos árboles de clasificación.**

***Index Terms*—Explicabilidad, Soft Computing, Árboles de decisión para Clasificación, Generación de Lenguaje Natural**

I. Introducción

les den explicaciones asociadas a las decisiones tomadas por los sistemas inteligentes que utilizan.

Desde un punto de vista técnico: ¿puede explicarnos la aplicación que tomó una decisión por qué tomó esa decisión y no otra? Para esto, hay básicamente dos opciones [5]: (1) el sistema inteligente está construido siguiendo un modelo interpretable (también llamado de caja blanca) que un operario experto puede analizar y entender a fin de elaborar una explicación; o (2) el sistema está construido siguiendo un modelo explicable que genera explicaciones por sí mismo. La DARPA planteó en 2016 las siguientes cuestiones técnicas [5]: ¿puede una máquina inteligente aprender de forma autónoma a explicar su comportamiento? ¿está preparada la generación

ExpliClas: Automatic Generation of Explanations in Natural Language for Weka Classifiers

Jose M. Alonso, A. Bugarín
Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS),
Universidade de Santiago de Compostela,
Campus Vida, E-15782, Santiago de Compostela, Spain
Email: {josemaria.alonso.moral, alberto.bugarin.diz}@usc.es

***Abstract*—ExpliClas is a web service aimed at providing users with multimodal (textual + graphical) explanations related to Weka classifiers. In ExpliClas, two types of explanations are automatically generated. On the one hand, global explanations pay attention to the behavior of the classifier as a whole, i.e., they refer to a list of structural properties (number of classes, features, etc.) along with quality indicators such as accuracy or confusion matrix. On the other hand, local explanations go in depth with how the classifier deals with single instances. Current version of ExpliClas already explains classifications made by three different decision tree Weka implementations (J48, RepTree, and RandomTree) and one fuzzy algorithm (FURIA). In this paper, we describe ExpliClas in detail and illustrate its use with the Bank telemarketing dataset.**

***Index Terms*—Explainable Artificial Intelligence, Natural Language Generation, Decision Trees, Fuzzy Unordered Rule Induction Algorithm, Open Source Software, Weka**

I. Introduction

We are living in a global technological world where everyone and everything is connected. In the data age, the main challenge is in the efficient and effective exploitation of the huge amount of data generated every second. Data scientists are required to extract valuable knowledge from the given data.

universities and R+D centers in USA are contributing to look for a new generation of XAI systems.

In addition, a new European General Data Protection Regulation (GDPR) was approved in 2016 and became effective in May 2018 [4]. The GDPR remarks the right to explanation of European citizens regarding decisions taken by automatic (intelligent) systems. Moreover, in June 2018, CLAIRE[1], an attempt to create a network of excellence in AI with the most well-recognized European universities and R+D centers, emphasized in its European human-centric vision for AI the need of building trustworthy AI that is beneficial to people through fairness, transparency, accountability and explainability.

It is worth noting that even though XAI systems are coming, there is still a lack of experts in this emerging research field. Moreover, new software tools are demanded with the aim of translating breakthrough knowledge and ideas into products and services for economic and social benefit [5].

In this paper, we introduce ExpliClas, i.e., a novel open source web service aimed at providing users with multimodal (textual + graphical) explanations related to AI classifiers. As far as we know, there is not any other similar web service in the world. ExpliClas can be seen as a dynamic dashboard

Towards Explaining Autonomy with Verbalised Decision Tree States

Konstantinos Gavriilidis[1], Andrea Munafo[2], Helen Hastie[1], Conlan Cesar[3], Michael DeFilippo[3], Michael R. Benjamin[3]
[1]*Edinburgh Centre for Robotics, Heriot-Watt University, Edinburgh, UK {kg47, H.Hastie}@hw.ac.uk*
[2]*SeeByte Ltd., Edinburgh, UK, andrea.munafo@seebyte.com*
[3]*Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, {conlanc, mikedefm, mikerb}@csail.mit.edu*

.13985v1 [cs.RO] 28 Sep 2022

I. Introduction

The development of new AUV technology increased the range of tasks that AUVs can tackle and the length of their operations. AUVs are today able to handle highly complex operations. However, these missions do not fit easily into the traditional method of defining a mission as a series of pre-planned waypoints because it is not possible to know, in advance, everything that might occur during the mission. This results in a gap between operator's expectations and actual operational performance. This then can create a diminished level of trust between the operators and AUVs, which can in turn result in unnecessary mission interruptions.

In behavioural autonomy [1]–[3], multiple behaviours are available to allow the robot to adapt to any circumstance and complete a mission. In Figure 1, for example, a simple sequence of behaviours is shown. In this case, the robot initially uses a *Survey* behaviour to explore an area, and once the objective *Survey1* is complete, a *Transit* behaviour is triggered to move to the next waypoint. During the transit, the vehicle could also trigger a *GPS* behaviour to obtain an GPS fix and update its position. This would temporarily interrupt its

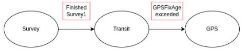


Fig. 1. Illustration of a behaviour chain together with the operational reason behind each activation.

autonomy from the decision points and the resulting executed actions applying *Knowledge Distillation* [6].

Knowledge distillation [6] makes it possible to interpret deterministic autonomous agents by building an equivalent representation, in our case, a distilled decision tree. The decision tree acts as a mediator retrieving the vehicle state in real time, and based on that, generates the corresponding state-actions tree traversals that match the autonomy decision-making process with the highest probability.

Finally, to present the explanations to the operators in a more natural way, the output of the distilled decision tree is combined with natural language explanations [7] and reported to the operators as sentences (see Figure 2). For this reason, an additional step known as *Concept2Text Generation* [8] is

Explaining Tree Model Decisions in Natural Language for Network Intrusion Detection (Ziems et al., NeurIPS, 2023)

Explaining Tree Model Decisions in Natural Language for Network Intrusion Detection

Noah Ziems, Gang Liu, John Flanagan, Meng Jiang
University of Notre Dame
{nziems2, gliu7, jflanag5, mjiang2}@nd.edu

Abstract

Network intrusion detection (NID) systems which leverage machine learning have been shown to have strong performance in practice when used to detect malicious network traffic. Decision trees in particular offer a strong balance between performance and simplicity, but require users of NID systems to have background knowledge in machine learning to interpret. In addition, they are unable to provide additional outside information as to why certain features may be important for classification.

In this work, we explore the use of large language models (LLMs) to provide explanations and additional background knowledge for decision tree NID systems. Further, we introduce a new human evaluation framework for decision tree explanations, which leverages automatically generated quiz questions that measure human evaluators' understanding of decision tree inference. Finally, we show LLM generated decision tree explanations correlate highly with human ratings of readability, quality, and use of background knowledge while simultaneously providing better understanding of decision boundaries.

1 Introduction

*“**LLM-based decision tree explanations (LLM-DTE)** […] leverage LLMs which have been autoregressively pretrained on large corpora spanning a wide range of topics.*

"LLM-based decision tree explanations (LLM-DTE) [...] leverage LLMs which have been autoregressively pretrained on large corpora spanning a wide range of topics.

*[...] **templates are used to convert important data** from the path and the decision tree **into a text format**.*

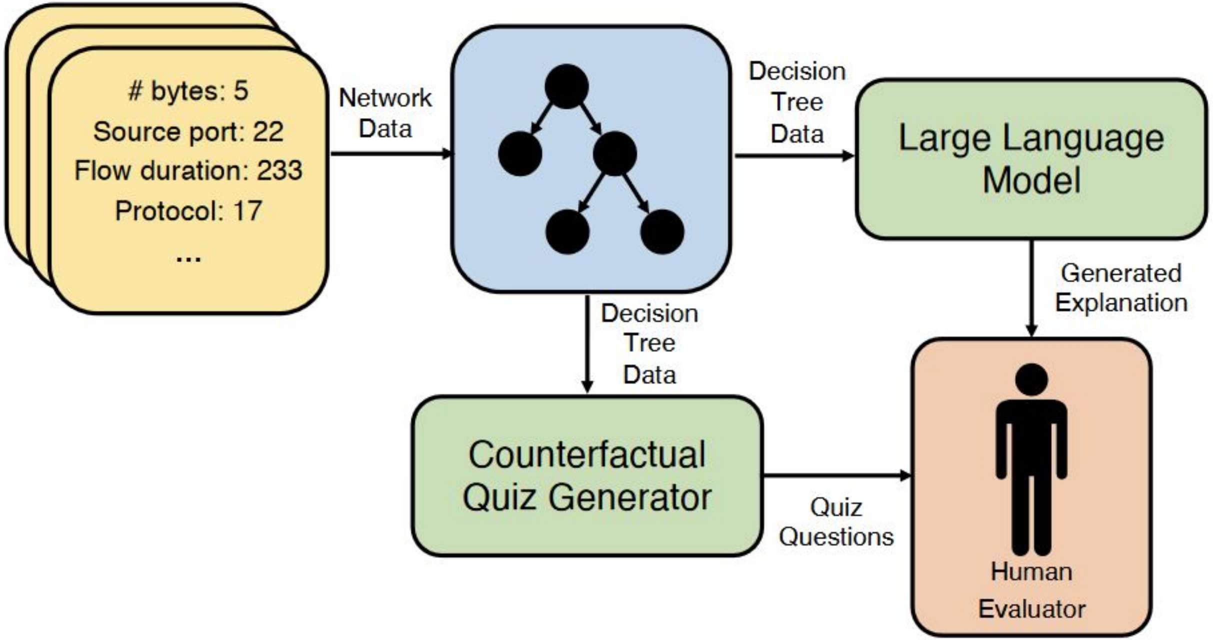
*[...] the filled-in **template is provided as a prompt to a LLM** which follows the prompt instructions to generate a natural language decision tree explanation."*

*“To construct our LLM prompt, we first **describe the task** of network intrusion detection and **provide feature descriptions** along with a **string-based representation of the trained decision tree**.*

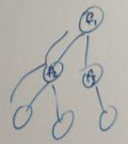
*Similar to our rule-based explanation, we add **a sentence for each node traversed in path** [...].*

*The **predicted label** from the decision tree **is provided** and emphasized.*

*Finally, the prompt is ended with **instructions to describe in simple terms** why the decision tree came to its conclusion.”*


bytes: 5
Source port: 22
Flow duration: 233
Protocol: 17
...
Network Data
Decision Tree Data
Large Language Model
Generated Explanation
Decision Tree Data
Counterfactual Quiz Generator
Quiz Questions
Human Evaluator

```python
def get_hydrated_prompt(example, orig_feat_names, feat_names, clf, df_orig, cat_columns, label,
prompt_id='a'):
      feature_desc = get_feature_descs(orig_feat_names)
      tree_text = export_text(clf, feature_names=feat_names)
      path_str, relevant_feature_str = print_tree_path(example, clf.tree_, feat_names, df_orig,
cat_columns)
      prompt = f"""Suppose a dataset for network intrusion detection has the following features:
      {feature_desc}
      The labels are Attack and Benign.
      The following decision tree was build using the above features:
      {tree_text}
      A new test example has the following relevant features:
      {relevant_feature_str}
      The new test example took the following path through the tree:
      {path_str}
      Using inferred background knowledge of the features and network traffic, explain in simple
terms why the decision tree came to the conclusion that the given example is {label}.
      Do not refer to the underlying mechanics of the decision tree in any way, and only refer to
the features using natural language. Please refer to the feature values in context using
parenthesis.
      """
      return prompt
```



⇓

tree_text

feature_names

⇓

feature_descriptions

tree_path

input →

"Chain-of-thoughts"

Let's think step-by-step question

"Suppose a dataset for...

{feature_descriptions}

The following DT was build... → LLM

ChatGPT-4

{tree_text}

The new test example ...

{tree_path}

⋮

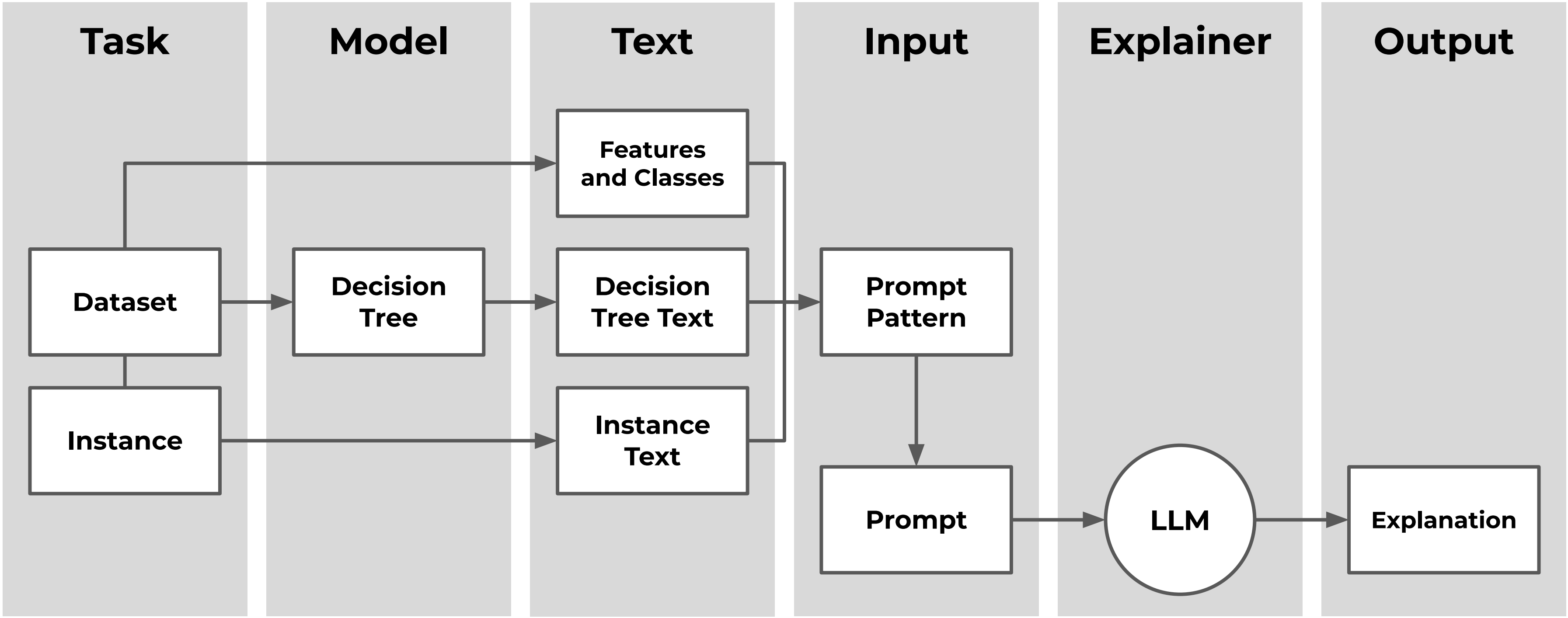
Model

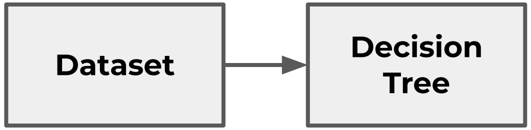
Domain

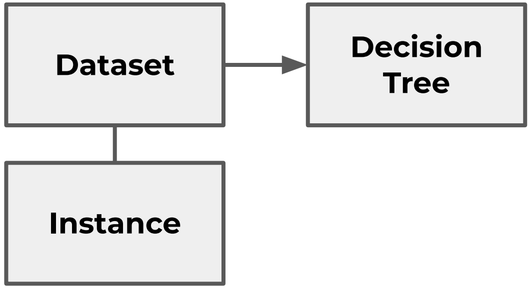
LLM

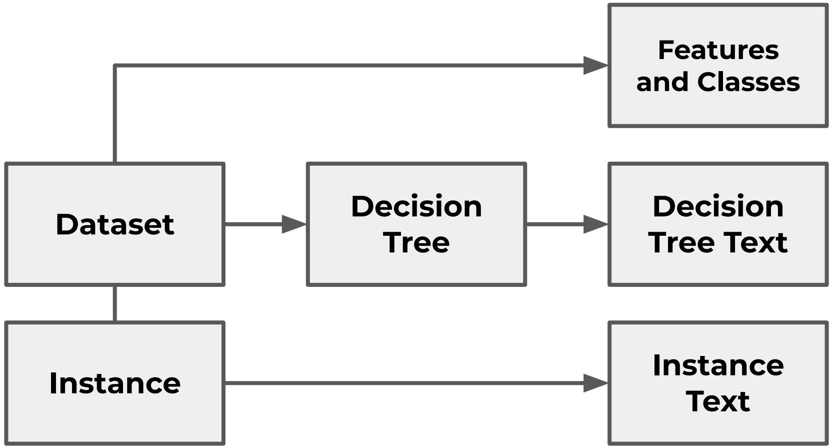
functions

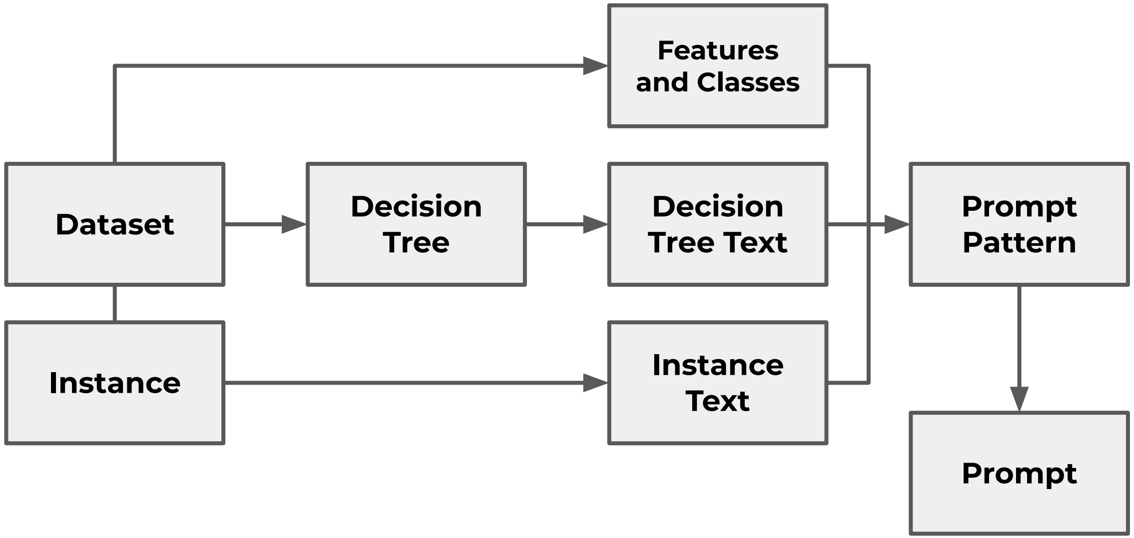
1. model instance to description
2. list feature descriptions
3. model behaviour (given a model instance and a dataset instance produces a description of the behaviour)


Task
Model
Text
Input
Explainer
Output
Features and Classes
Dataset
Decision Tree
Decision Tree Text
Prompt Pattern
Instance
Instance Text
Prompt
LLM
Explanation

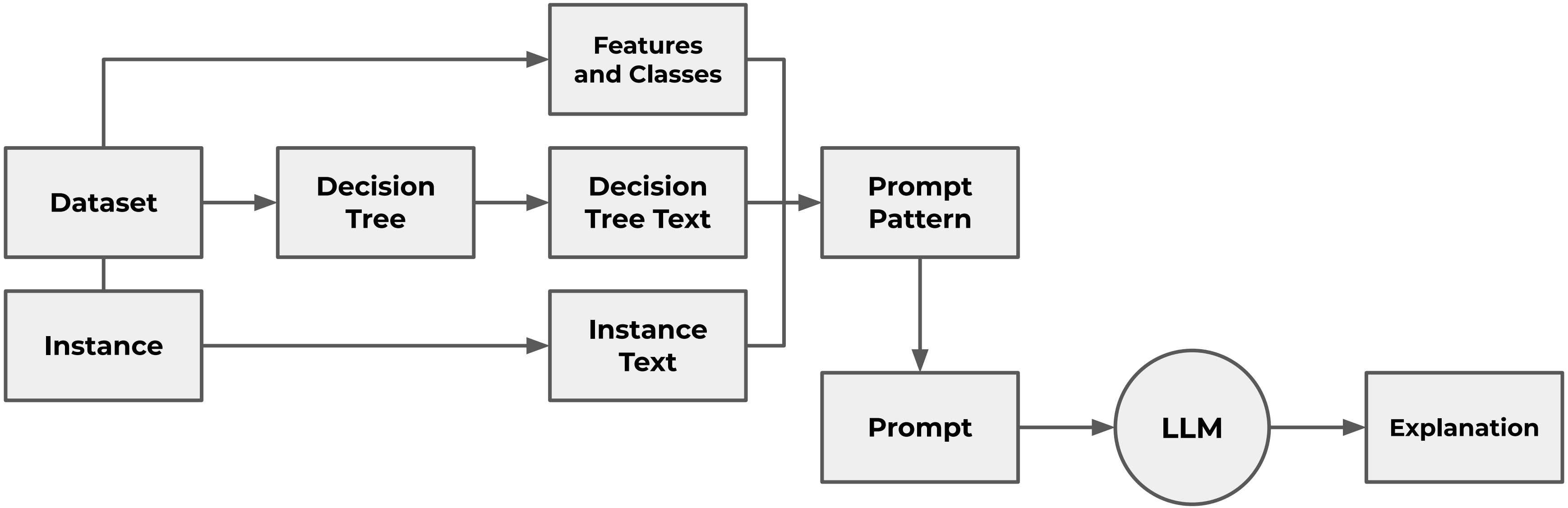

Dataset
Decision Tree

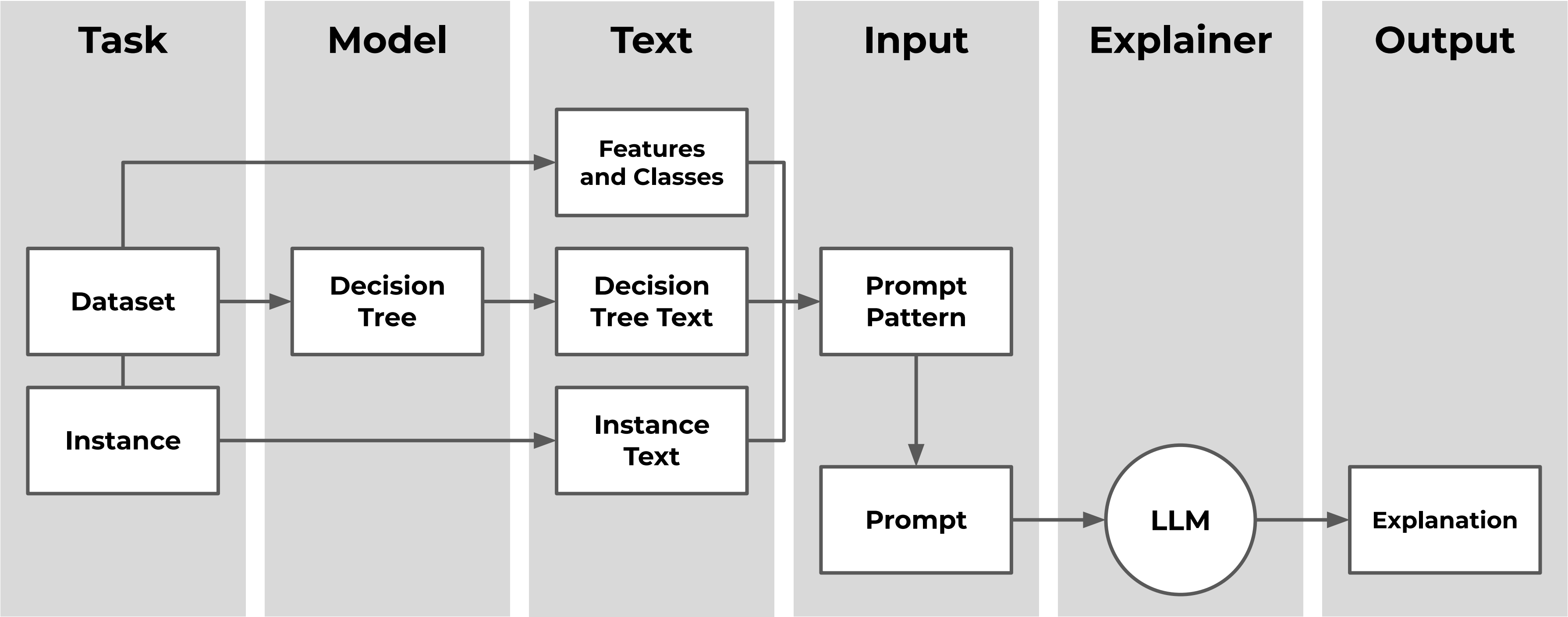

Dataset
Decision Tree
Instance


Features
and Classes
Dataset
Decision
Tree
Decision
Tree Text
Instance
Instance
Text



Features and Classes
Dataset
Decision Tree
Decision Tree Text
Prompt Pattern
Instance
Instance Text
Prompt


Features
and Classes
Dataset
Decision
Tree
Decision
Tree Text
Prompt
Pattern
Instance
Instance
Text
Prompt
LLM
Explanation


Task
Model
Text
Input
Explainer
Output
Features and Classes
Dataset
Decision Tree
Decision Tree Text
Prompt Pattern
Instance
Instance Text
Prompt
LLM
Explanation

Experiments

Prompt Engineering

- Direct Question
- Demonstration
- Instructions

Text Representation

- Incorrect Classification
- Tree Text
- External Demonstration
- Demonstration from Different Task

Prompt Pattern - Direct Question

Consider a dataset that has the following features: **{features_names}**. Each instance can be classified into one of the following classes: **{classes}**. A decision tree was trained on the dataset and the following tree was obtained:

{tree_text}

An instance has features: **{instance}**.

Please explain in simple terms why the decision tree concluded that the given example is **{predicted_class}** with a confidence of **{confidence_value}**.

<table>
<tr><td>Features</td><td><code>sepal length (cm), sepal width (cm), petal length (cm), and petal width (cm)</code></td></tr>
<tr><td>Classes</td><td><code>setosa, versicolor, and virginica</code></td></tr>
<tr><td>Decision Tree</td><td><pre>
|--- petal length (cm) <= 2.45
| |--- class: 0
|--- petal length (cm) > 2.45
| |--- petal width (cm) <= 1.55
| | |--- petal length (cm) <= 4.95
| | | |--- class: 1
| | |--- petal length (cm) > 4.95
| | | |--- class: 2
| |--- petal width (cm) > 1.55
| | |--- petal width (cm) <= 1.70
| | | |--- class: 1
| | |--- petal width (cm) > 1.70
| | | |--- class: 2
</pre></td></tr>
<tr><td>Instance features</td><td><code>sepal length (cm) = 7.3, sepal width (cm) = 2.9, petal length (cm) = 6.3, and petal width (cm) = 1.8</code></td></tr>
<tr><td>Instance class</td><td><code>virginica</code></td></tr>
<tr><td>Confidence</td><td><code>97.06 %.</code></td></tr>
</table>

Prompt - Direct Question

Consider a dataset that has the following features:**sepal length (cm), sepal width (cm), petal length (cm), and petal width (cm).** Each instance can be classified into one of the following classes: **setosa, versicolor, and virginica.** A decision tree was trained on the dataset and the following tree was obtained:

```
|--- petal length (cm) <= 2.45
|   |--- class: 0
|--- petal length (cm) >  2.45
|   |--- petal width (cm) <= 1.55
|   |   |--- petal length (cm) <= 4.95
|   |   |   |--- class: 1
|   |   |--- petal length (cm) >  4.95
|   |   |   |--- class: 2
|   |--- petal width (cm) >  1.55
|   |   |--- petal width (cm) <= 1.70
|   |   |   |--- class: 1
|   |   |--- petal width (cm) >  1.70
|   |   |   |--- class: 2
```

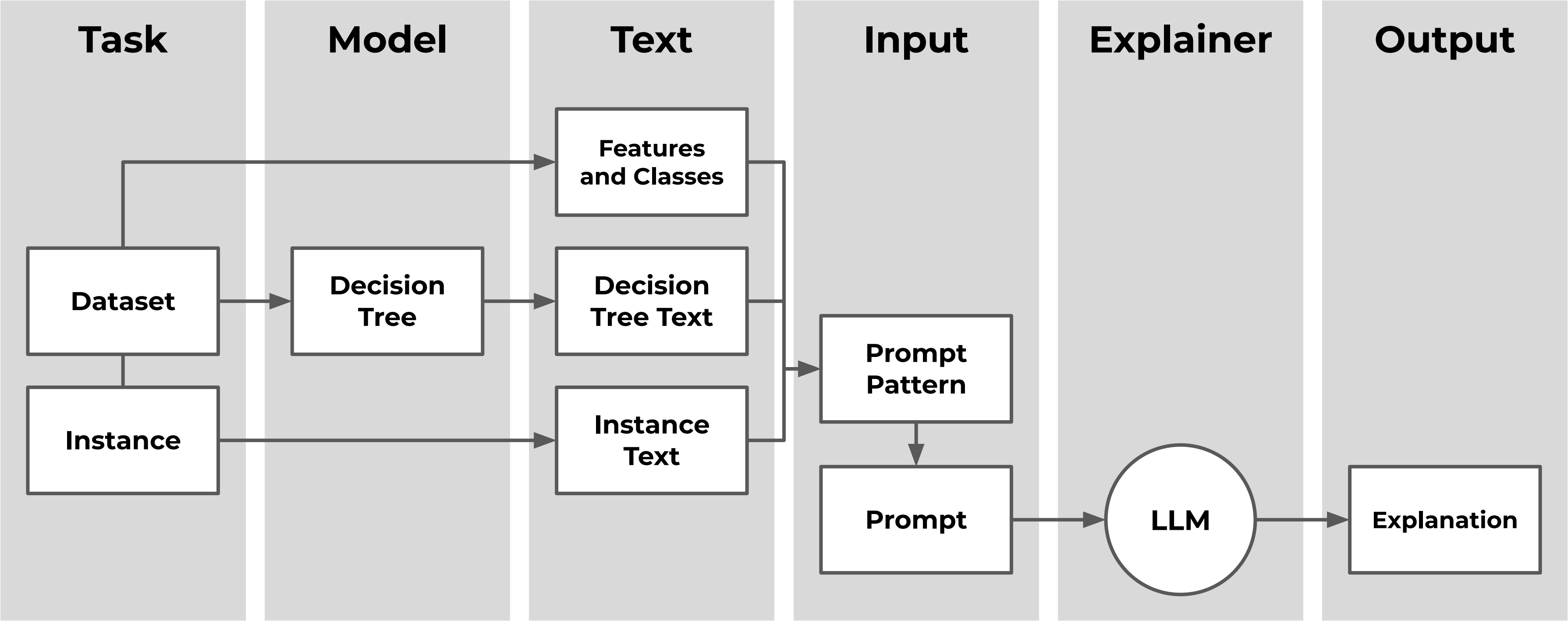
An instance has features: **sepal length (cm) = 7.3, sepal width (cm) = 2.9, petal length (cm) = 6.3, and petal width (cm) = 1.8.**
Please explain in simple terms why the decision tree concluded that the given example is **virginica** with a confidence of **97.06 %.**

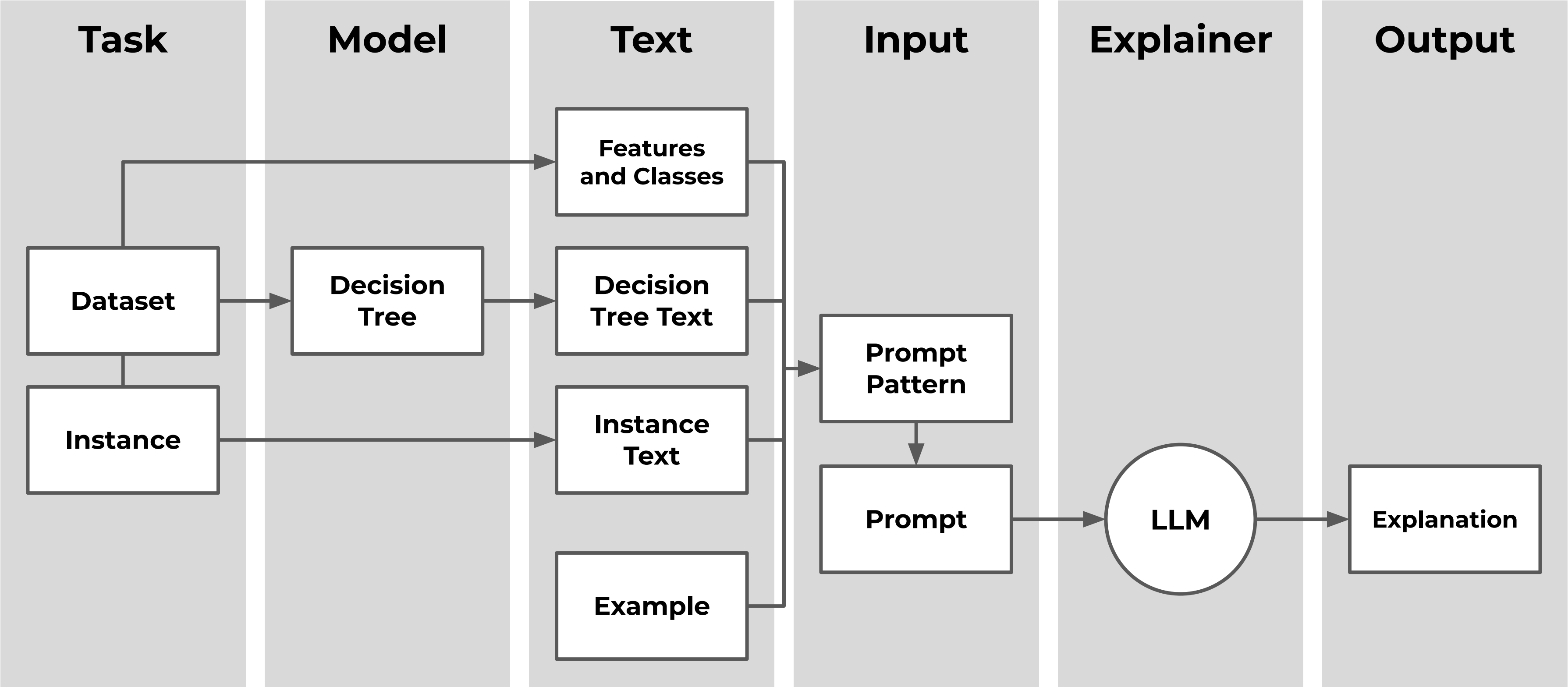
Response - Direct Question

The decision tree concluded that the given example is virginica with a confidence of 97.06% because it followed a series of rules based on the features of the instance.

First, it checked if the petal length was less than or equal to 2.45 cm, which was not the case for the given example. Then, it checked if the petal width was less than or equal to 1.55 cm, which was also not the case. Finally, it checked if the petal width was greater than 1.70 cm, which was true for the given example.

Based on these rules, the decision tree determined that the given example belongs to the class virginica with a high confidence of 97.06%.


Task
Model
Text
Input
Explainer
Output
Features and Classes
Dataset
Decision Tree
Decision Tree Text
Prompt Pattern
Instance
Instance Text
Prompt
LLM
Explanation


Task
Model
Text
Input
Explainer
Output
Features and Classes
Dataset
Decision Tree
Decision Tree Text
Prompt Pattern
Instance
Instance Text
Prompt
LLM
Explanation
Example

Prompt Pattern - Demonstration

Consider a dataset that has the following features: **{features_names}**. Each instance can be classified into one of the following classes: **{classes}**. A decision tree was trained on the dataset and the following tree was obtained:

{tree_text}

{demonstration}

An instance has features: **{instance}**.

Please explain in similar terms why the decision tree concluded that the given example is **{predicted_class}** with a confidence of **{confidence_value}**.

Features	sepal length (cm), sepal width (cm), petal length (cm), and petal width (cm)
Classes	setosa, versicolor, and virginica

Decision Tree	<pre>|--- petal length (cm) <= 2.45
	| |--- class: 0
	|--- petal length (cm) > 2.45
	| |--- petal width (cm) <= 1.55
	| | |--- petal length (cm) <= 4.95
	| | | |--- class: 1
	| | |--- petal length (cm) > 4.95
	| | | |--- class: 2
	| |--- petal width (cm) > 1.55
	| | |--- petal width (cm) <= 1.70
	| | | |--- class: 1
	| | |--- petal width (cm) > 1.70
	| | | |--- class: 2</pre>

Demonstration	Given an instance of the iris dataset with features: sepal length (cm) = 3.6, sepal width (cm) = 0.8, petal length (cm) = 4.3, and petal width (cm) = 2.9, and a confidence value of 57.81 %, a **good explanation** for why the instance was classified as virginica is: **'By evaluating the feature values, it is possible to observe that both petal length and petal width are high. This means that by following the decision tree path, the instance should be classified as virginica, although the tree is not very confident in this result.'**
Instance	sepal length (cm) = 7.3, sepal width (cm) = 2.9, petal length (cm) = 6.3, and petal width (cm) = 1.8; virginica; 97.06 %

Prompt - Demonstration

Consider a dataset that has the following features: sepal length (cm), sepal width (cm), petal length (cm), and petal width (cm). Each instance can be classified into one of the following classes: setosa, versicolor, and virginica. A decision tree was trained on the dataset and the following tree was obtained:

```
|--- petal length (cm) <= 2.45
|   |--- class: 0
|--- petal length (cm) >  2.45
|   |--- petal width (cm) <= 1.55
|   |   |--- petal length (cm) <= 4.95
|   |   |   |--- class: 1
|   |   |--- petal length (cm) >  4.95
|   |   |   |--- class: 2
|   |--- petal width (cm) >  1.55
|   |   |--- petal width (cm) <= 1.70
|   |   |   |--- class: 1
|   |   |--- petal width (cm) >  1.70
|   |   |   |--- class: 2
```

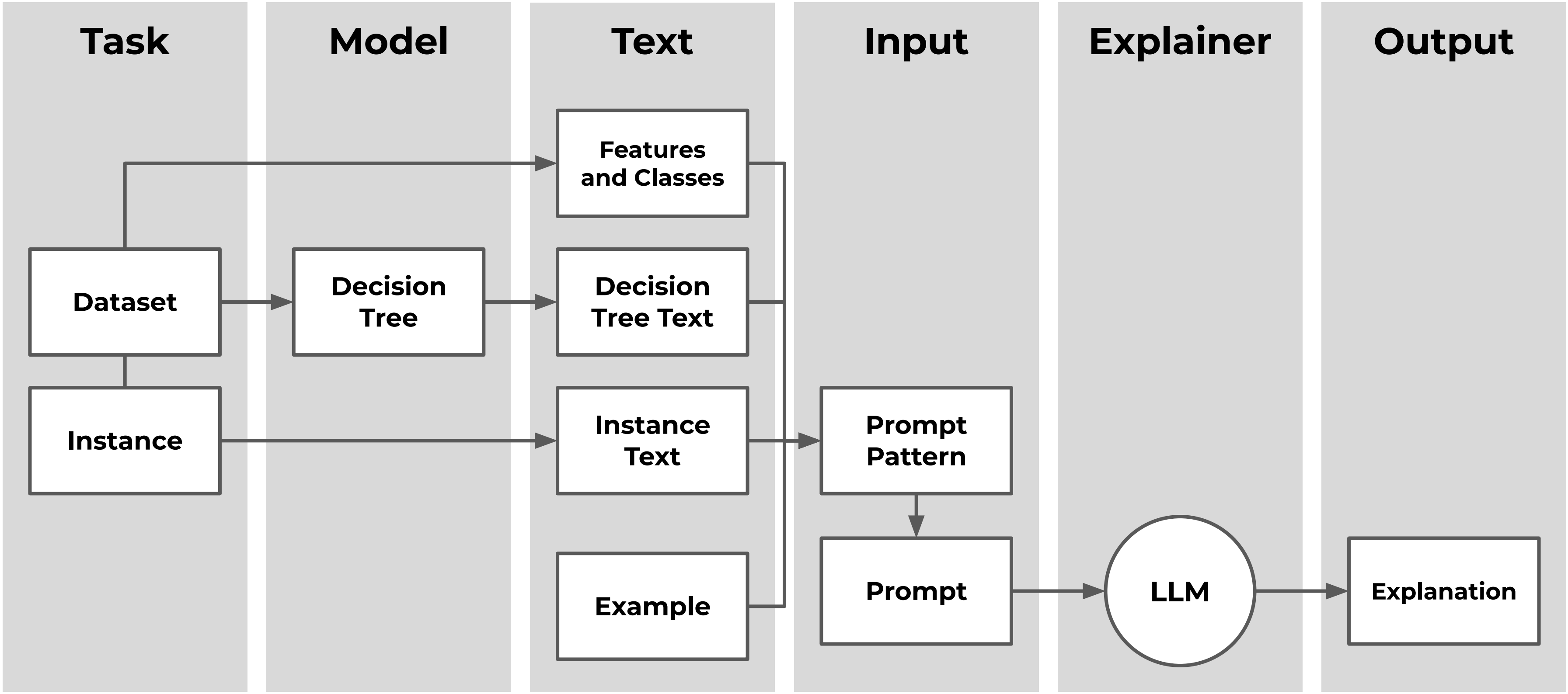
Given an instance of the iris dataset with features: sepal length (cm) = 3.6, sepal width (cm) = 0.8, petal length (cm) = 4.3, and petal width (cm) = 2.9, and a confidence value of 57.81 %, a good explanation for why the instance was classified as virginica is: 'By evaluating the feature values, it is possible to observe that both petal length and petal width are high. This means that by following the decision tree path, the instance should be classified as virginica, although the tree is not very confident in this result.'

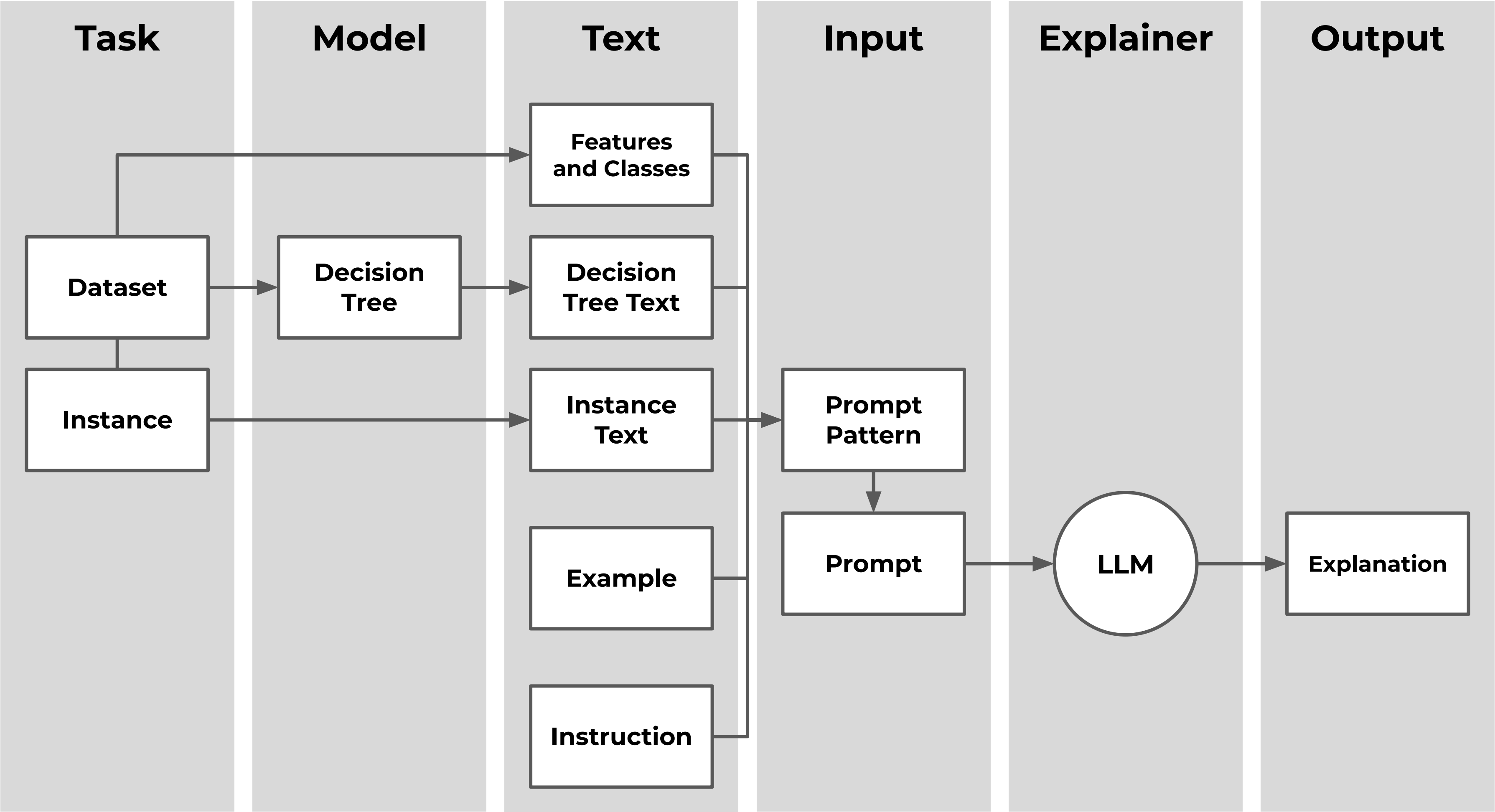
An instance has features: sepal length (cm) = 7.3, sepal width (cm) = 2.9, petal length (cm) = 6.3, and petal width (cm) = 1.8.

Please explain in **similar** terms why the decision tree concluded that the given example is virginica with a confidence of 97.06 %.

Response - Demonstration

The decision tree classified the given example as virginica with a confidence of 97.06% because it evaluated the feature values and found that both petal length and petal width are high. According to the decision tree path, when petal length is greater than 2.45 and petal width is greater than 1.55, the instance is classified as virginica. In this case, the high values of petal length (6.3 cm) and petal width (1.8 cm) led the decision tree to confidently classify the instance as virginica.


Task
Model
Text
Input
Explainer
Output
Features and Classes
Dataset
Decision Tree
Decision Tree Text
Instance
Instance Text
Prompt Pattern
Example
Prompt
LLM
Explanation


Task
Model
Text
Input
Explainer
Output
Features and Classes
Dataset
Decision Tree
Decision Tree Text
Instance
Instance Text
Prompt Pattern
Example
Prompt
LLM
Explanation
Instruction

Prompt Pattern - Instructions

Consider a dataset that has the following features: **{features_names}**. Each instance can be classified into one of the following classes: **{classes}**. A decision tree was trained on the dataset and the following tree was obtained:

{tree_text}

{demonstration}

An instance has features: **{instance}**.

{instructions}

Please explain in similar terms why the decision tree concluded that the given example is **{predicted_class}** with a confidence of **{confidence_value}**.

Features	sepal length (cm), sepal width (cm), petal length (cm), and petal width (cm)
Classes	setosa, versicolor, and virginica

Decision Tree	`|--- petal length (cm) <= 2.45`
	`|   |--- class: 0`
	`|--- petal length (cm) >  2.45`
	`|   |--- petal width (cm) <= 1.55`
	`|   |   |--- petal length (cm) <= 4.95`
	`|   |   |   |--- class: 1`
	`|   |   |--- petal length (cm) >  4.95`
	`|   |   |   |--- class: 2`
	`|   |--- petal width (cm) >  1.55`
	`|   |   |--- petal width (cm) <= 1.70`
	`|   |   |   |--- class: 1`
	`|   |   |--- petal width (cm) >  1.70`
	`|   |   |   |--- class: 2`

Demonstration	Given an instance of the iris dataset with features: sepal length (cm) = 3.6, sepal width (cm) = 0.8, petal length (cm) = 4.3, and petal width (cm) = 2.9, and a confidence value of 57.81 %, a good explanation for why the instance was classified as virginica is: 'By evaluating the feature values, it is possible to observe that both petal length and petal width are high. This means that by following the decision tree path, the instance should be classified as virginica, although the tree is not very confident in this result.'
Instance	sepal length (cm) = 7.3, sepal width (cm) = 2.9, petal length (cm) = 6.3, and petal width (cm) = 1.8
Instructions	For answering the following question, **do not refer to the underlying mechanics of the decision tree** in any way, and only refer to the features using natural language. All the relevant features must be mentioned in the answer, but features that were not used by the tree should be ignored. Moreover, **do not use any technical jargon or numerical values in the response and prefer to user terms like 'high' and 'low'**.
Question	Please explain in similar terms why the decision tree concluded that the given example is virginica with a confidence of 97.06 %.

Prompt - Instructions

Consider a dataset that has the following features: sepal length (cm), sepal width (cm), petal length (cm), and petal width (cm). Each instance can be classified into one of the following classes: setosa, versicolor, and virginica. A decision tree was trained on the dataset and the following tree was obtained:

```
|--- petal length (cm) <= 2.45
|   |--- class: 0
|--- petal length (cm) >  2.45
|   |--- petal width (cm) <= 1.55
|   |   |--- petal length (cm) <= 4.95
|   |   |   |--- class: 1
|   |   |--- petal length (cm) >  4.95
|   |   |   |--- class: 2
|   |--- petal width (cm) >  1.55
|   |   |--- petal width (cm) <= 1.70
|   |   |   |--- class: 1
|   |   |--- petal width (cm) >  1.70
|   |   |   |--- class: 2
```

Given an instance of the iris dataset with features: sepal length (cm) = 3.6, sepal width (cm) = 0.8, petal length (cm) = 4.3, and petal width (cm) = 2.9, and a confidence value of 57.81 %, a good explanation for why the instance was classified as virginica is: 'By evaluating the feature values, it is possible to observe that both petal length and petal width are high. This means that by following the decision tree path, the instance should be classified as virginica, although the tree is not very confident in this result.'

An instance has features: sepal length (cm) = 7.3, sepal width (cm) = 2.9, petal length (cm) = 6.3, and petal width (cm) = 1.8.

For answering the following question, do not refer to the underlying mechanics of the decision tree in any way, and only refer to the features using natural language. All the relevant features must be mentioned in the answer, but features that were not used by the tree should be ignored. Moreover, do not use any technical jargon or numerical values in the response and prefer to user terms like 'high' and 'low'.

Please explain in similar terms why the decision tree concluded that the given example is virginica with a confidence of 97.06 %.

Response - Instructions

The decision tree concluded that the given example is virginica with a confidence of 97.06% because the **petal length is high** and the **petal width is also high.** This combination of high petal length and petal width led the decision tree to classify the instance as virginica with **high confidence.**

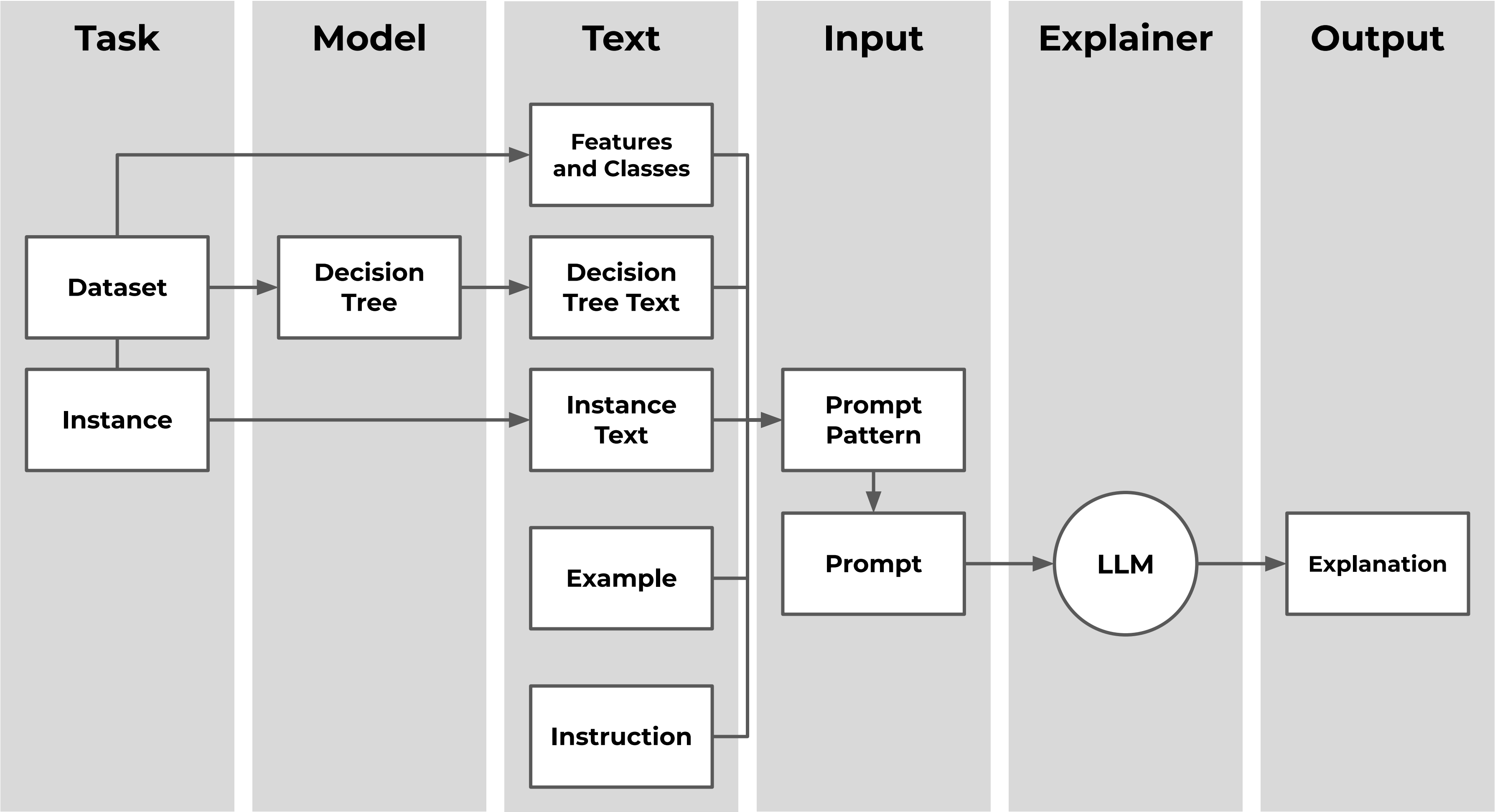
Experiments

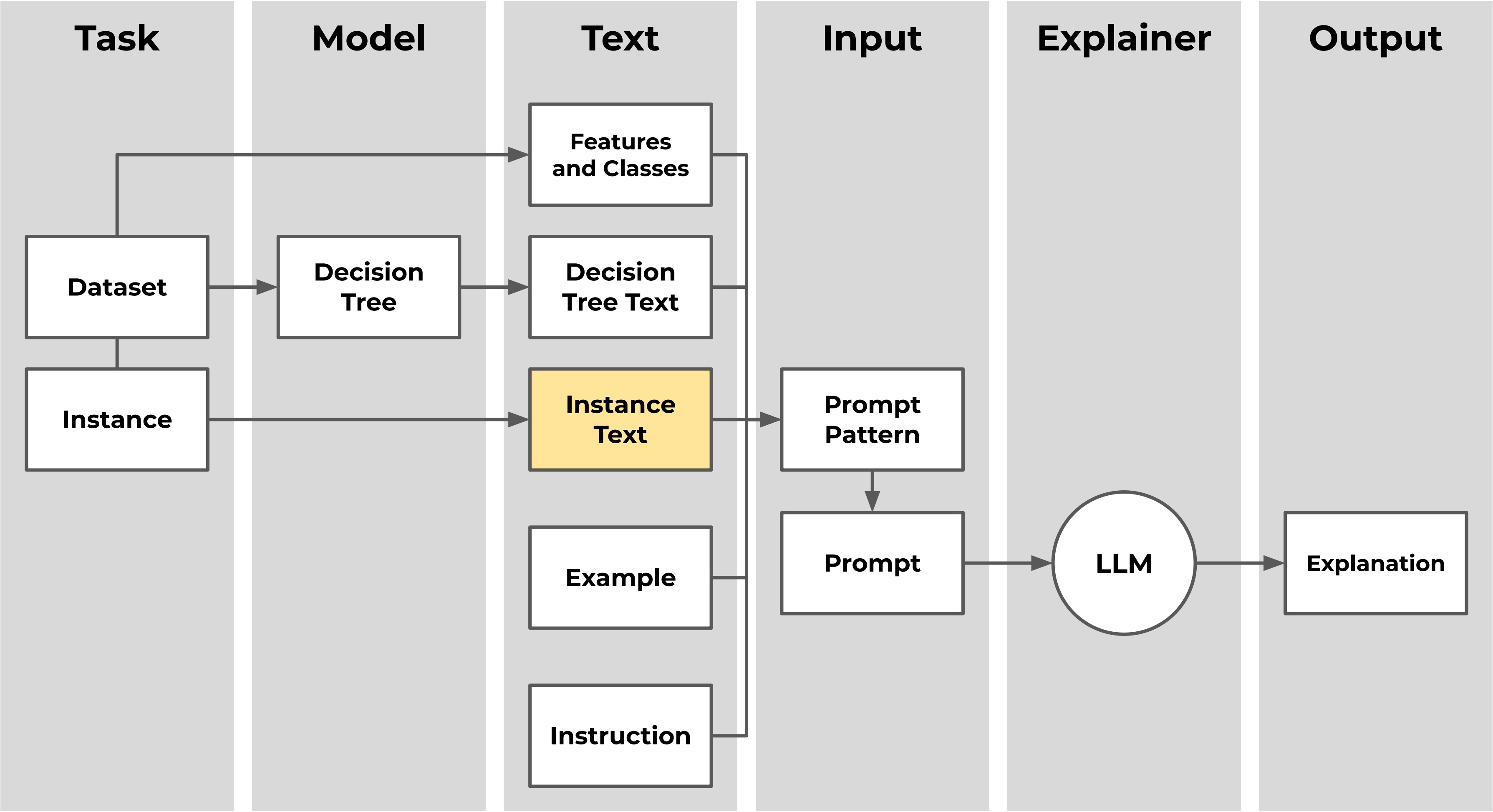
Prompt Engineering

- Direct Question
- Demonstration
- Instructions

Text Representation

- **Incorrect Classification**
- **Tree Text**
- **External Demonstration**
- **Demonstration from Different Task**


Task
Model
Text
Input
Explainer
Output
Features and Classes
Dataset
Decision Tree
Decision Tree Text
Instance
Instance Text
Prompt Pattern
Example
Prompt
LLM
Explanation
Instruction


Task
Model
Text
Input
Explainer
Output
Features and Classes
Dataset
Decision Tree
Decision Tree Text
Instance
Instance Text
Prompt Pattern
Example
Prompt
LLM
Explanation
Instruction

Incorrect Classification

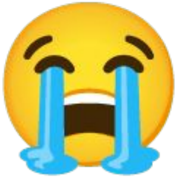
Features	Correct Class	Predicted Class
sepal length (cm) = 3.6 sepal width (cm) = 0.8 petal length (cm) = 4.3 petal width (cm) = 2.9	Virginica	Versicolor

Incorrect Classification - Response

The decision tree concluded that the given example is versicolor with a confidence of 97.06% because the **petal length is high** and the **petal width is low**. This combination of feature values led the tree to classify the instance as versicolor with high confidence.

Incorrect Classification - Response

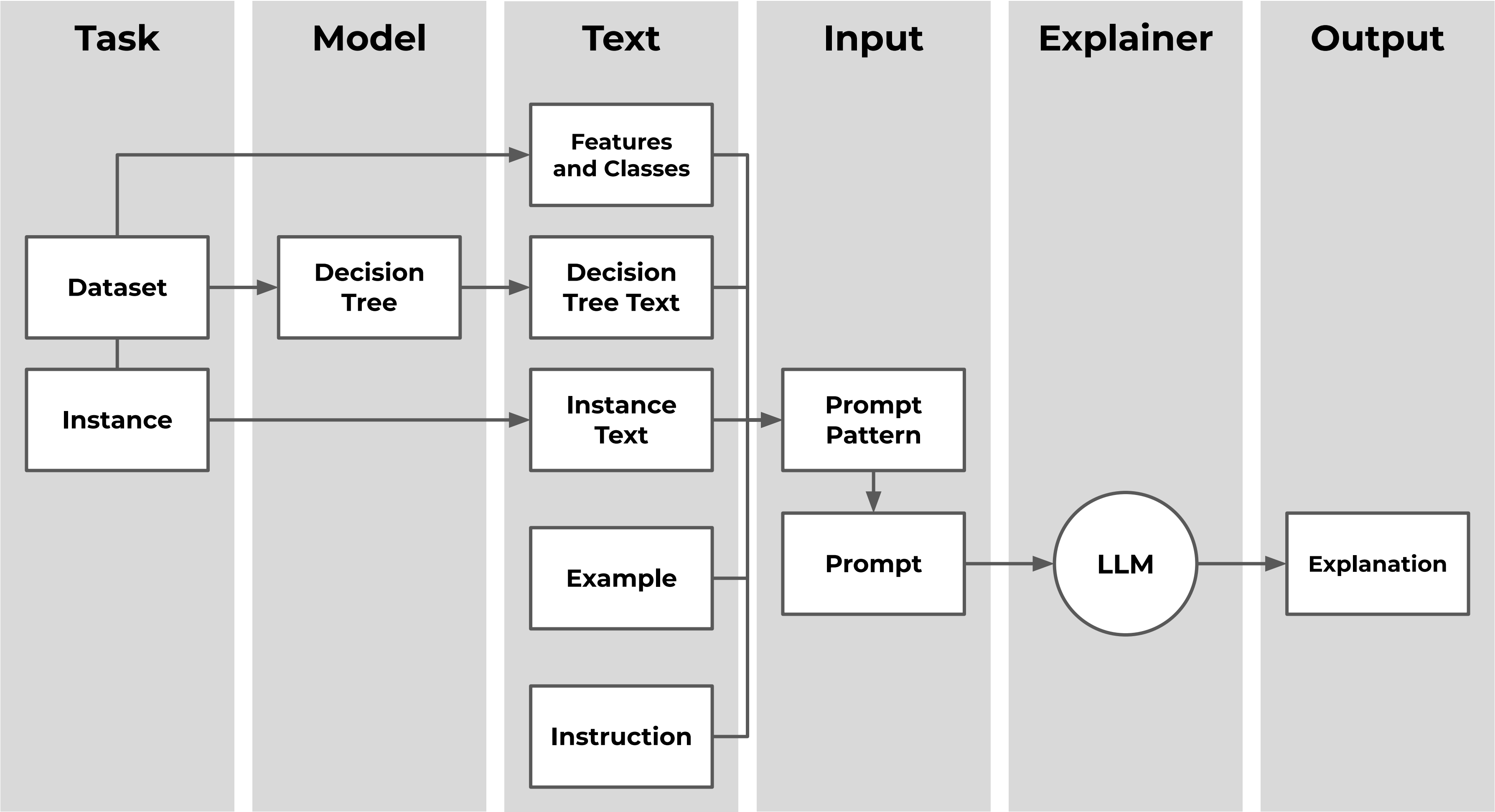
The decision tree concluded that the given example is versicolor with a confidence of 97.06% because the **petal length is high** and the **petal width is low.** This combination of feature values led the tree to classify the instance as versicolor with high confidence.



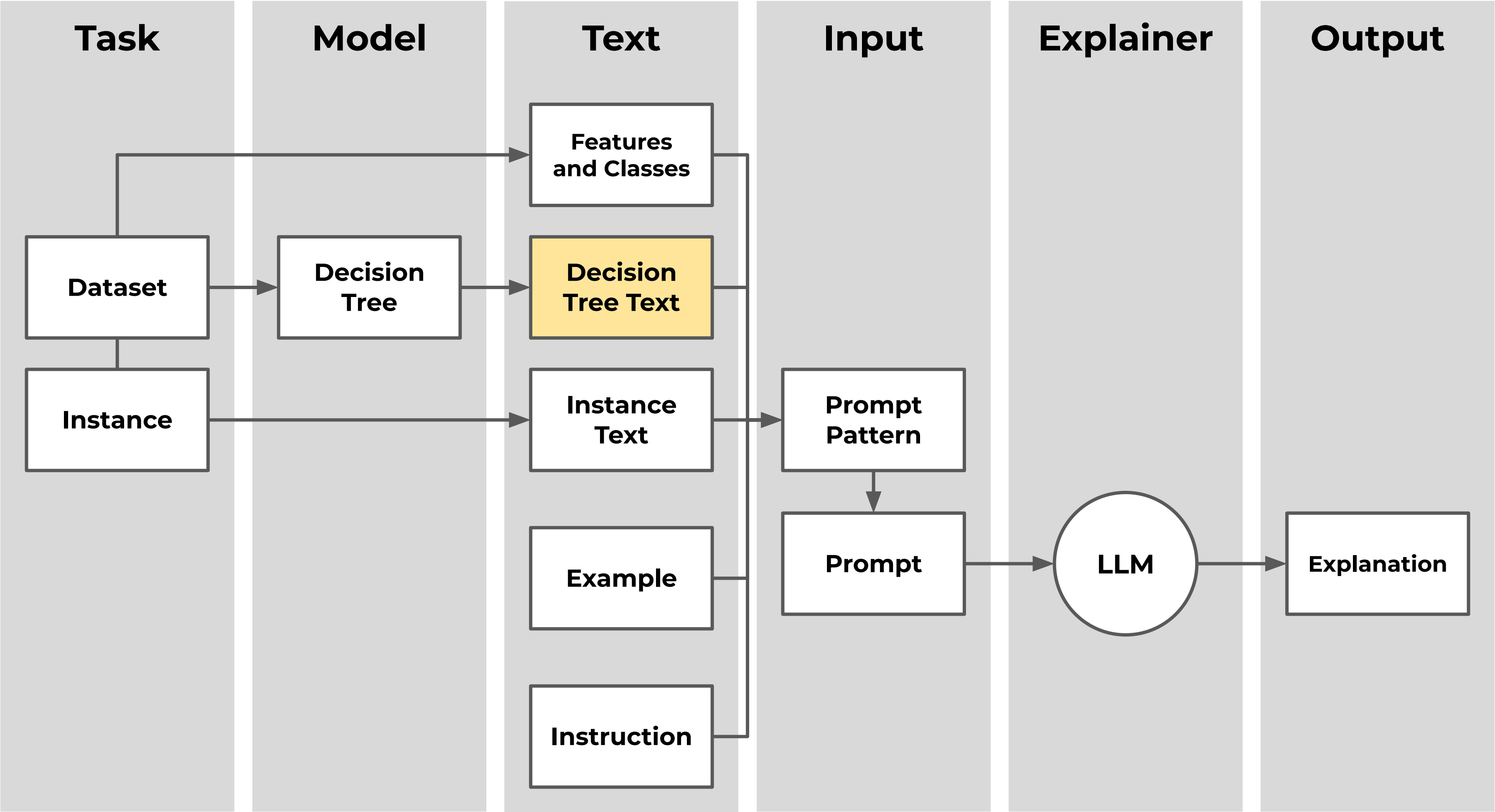
Correct prediction:

*"because the **petal length is high** and the **petal width is also high.**"*

Current instructions are not enough, because the tree doesn't check for correctness



Task
Model
Text
Input
Explainer
Output
Features and Classes
Dataset
Decision Tree
Decision Tree Text
Instance
Instance Text
Prompt Pattern
Prompt
LLM
Explanation
Example
Instruction


Task
Model
Text
Input
Explainer
Output
Features and Classes
Dataset
Decision Tree
Decision Tree Text
Instance
Instance Text
Prompt Pattern
Example
Prompt
LLM
Explanation
Instruction

Change Tree Representation

Original

```
|--- petal length (cm) <= 2.45
|   |--- class: 0
|--- petal length (cm) >  2.45
|   |--- petal width (cm) <= 1.55
|   |   |--- petal length (cm) <= 4.95
|   |   |   |--- class: 1
|   |   |--- petal length (cm) >  4.95
|   |   |   |--- class: 2
|   |--- petal width (cm) >  1.55
|   |   |--- petal width (cm) <= 1.70
|   |   |   |--- class: 1
|   |   |--- petal width (cm) >  1.70
|   |   |   |--- class: 2
```

New

```
Root Node:
- Feature: Petal Length
- Condition: If Petal Length ≤ 2.45 cm
Decision Node 1 (Left Child):
- Leaf Node: Setosa
Decision Node 2 (Right Child):
- Feature: Petal Width
- Condition: If Petal Width ≤ 1.55 cm
   Decision Node 3 (Left Child):
   - Feature: Petal Length
   - Condition: If Petal Length ≤ 4.95 cm
        - Leaf Node: Versicolor
   - Right Node: Virginica
   Decision Node 4 (Right Child):
   - Feature: Petal Width
   - Condition: If Petal Length ≤ 1.70 cm
        - Leaf Node: Versicolor
   - Right Node: Virginica
```

Change Tree Representation - Response

The decision tree classified the given example as virginica with a **high confidence** because the **petal length is high** and the **petal width is also high**. This combination of high petal length and petal width led the decision tree to conclude that the instance belongs to the virginica class.

Change Tree Representation - Response

The decision tree classified the given example as virginica with a high confidence **because the petal length is high and the petal width is also high. This combination of high petal length and petal width led the decision tree to** conclude that the instance belongs to the virginica class .

Original:

The decision tree concluded that the given example is virginica with a confidence of 97.06% **because the petal length is high and the petal width is also high. This combination of high petal length and petal width led the decision tree to** classify the instance as virginica with high confidence .

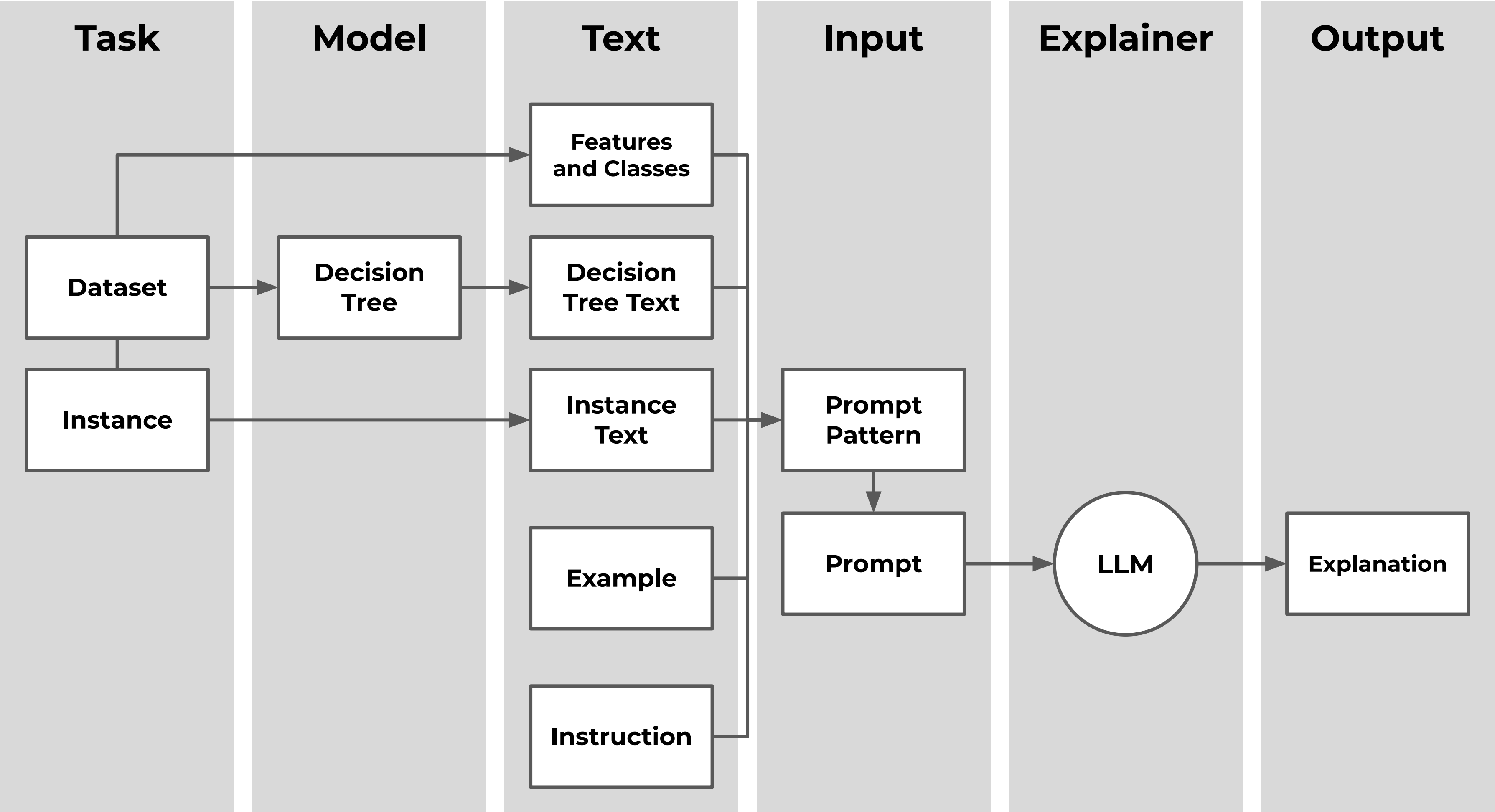
Change Tree Representation - Response

The decision tree classified the given example as virginica with a high confidence **because the petal length is high and the petal width is also high. This combination of high petal length and petal width led the decision tree to** conclude that the instance belongs to the virginica class .

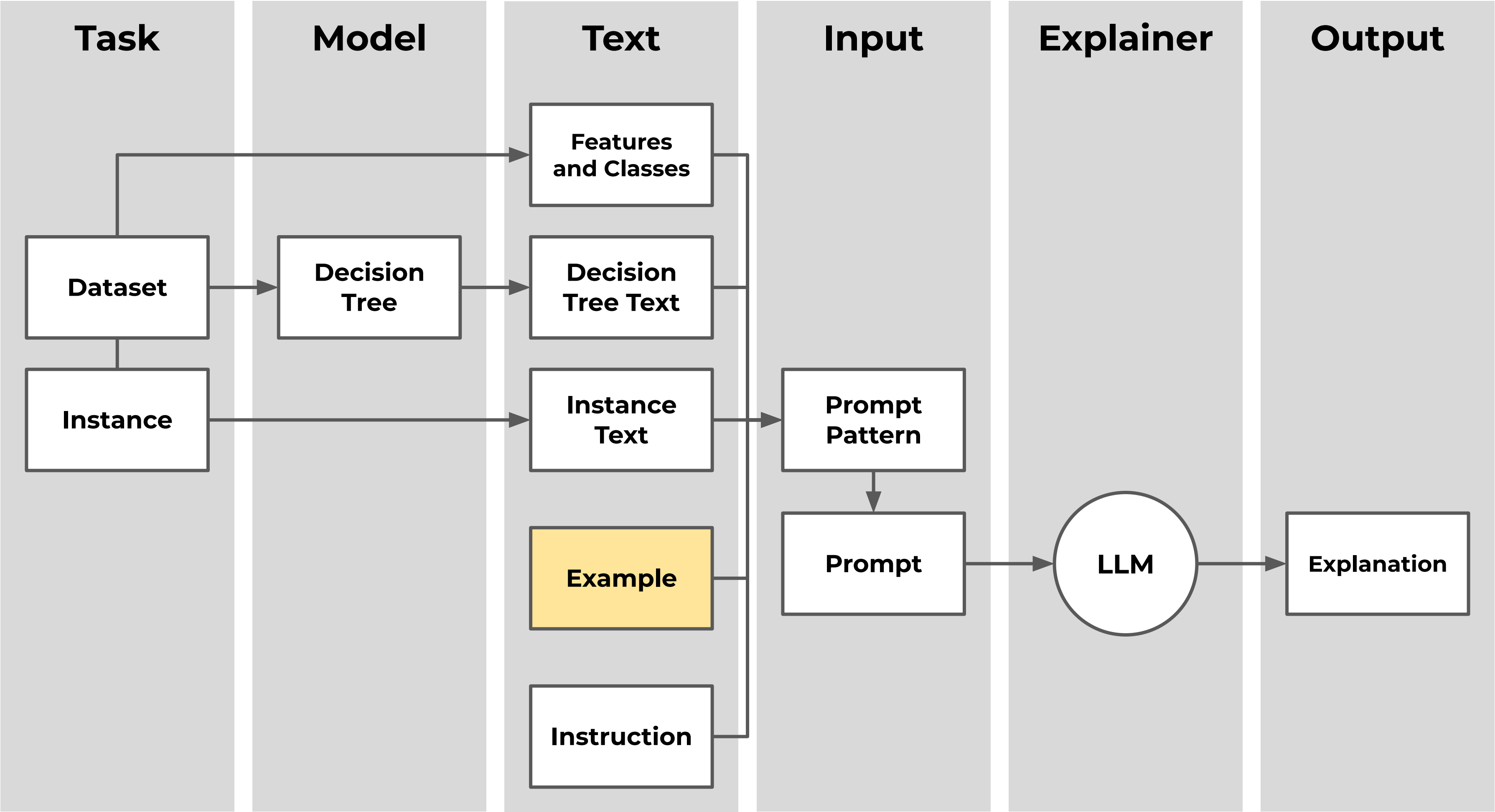
Original:

The decision tree concluded that the given example is virginica with a confidence of 97.06% **because the petal length is high and the petal width is also high. This combination of high petal length and petal width led the decision tree to** classify the instance as virginica with high confidence .

It seems that the explanation is agnostic to the textual representation of the tree



Task
Model
Text
Input
Explainer
Output
Features and Classes
Dataset
Decision Tree
Decision Tree Text
Instance
Instance Text
Prompt Pattern
Example
Prompt
LLM
Explanation
Instruction


Task
Model
Text
Input
Explainer
Output
Features and Classes
Dataset
Decision Tree
Decision Tree Text
Instance
Instance Text
Prompt Pattern
Example
Prompt
LLM
Explanation
Instruction

Generación Automática de Explicaciones en Lenguaje Natural para Árboles de Decisión de Clasificación

B. López-Trigo, Jose M. Alonso, A. Bugarín
Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS),
Universidade de Santiago de Compostela,

Generación Automática de Explicaciones en Lenguaje Natural para Árboles de Decisión de Clasificación

B. López-Trigo, Jose M. Alonso, A. Bugarín
Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS),
Universidade de Santiago de Compostela,

```
cases.
```

IV. Algunos Casos de Uso

Una vez descritos los elementos que componen cada explicación, veremos en esta sección dos ejemplos completos, con los que ilustraremos el funcionamiento de nuestra propuesta paso a paso. En ambos casos se aprenden clasificadores utilizando el algoritmo C4.5 [10], en la implementación disponible en Weka (J48) [16], [17]. Tanto los dos ejemplos mostrado (IRIS y FLAVIA), como otros disponibles, se pueden reproducir con el servicio Web ExpliClas [1] (Fig. 1).

IV-A. Conjunto de datos IRIS

El conjunto de datos IRIS (uno de los más conocidos del repositorio [18]) está formado por 150 instancias, 4 atributos numéricos y 3 clases. El árbol de clasificación generado por Weka (Fig. 2) está formado por 9 nodos totales, 5 de ellos nodos-hoja que deciden la clasificación y los 4 nodos restantes con las condiciones (comparaciones sobre los valores de los atributos) para decidir la clasificación. Se trata, por tanto, de un árbol simple que utilizaremos como primer ejemplo.

La explicación global generada en este caso es la siguiente:

```
There are 3 types of iris:
Setosa, Virginica and Versicolor.
This classifier is very reliable
because correctly classified
instances represent 96%.
```

La explicación local, para la instancia de la Fig. 3 (Sepal-Length: 5.6, Sepal-Width: 3, Petal-Length: 4.1, Petal-Width: 1.3) es la siguiente:

```
Iris is type Virginica because
its petal-length and petal-width
are medium.
```

En este caso, la explicación consiste en indicar los valores lingüísticos correspondientes a los valores numéricos de los atributos que han dado lugar a la clasificación, tal y como se detalla en la figura.

Sin embargo, si tomamos una instancia cuyos valores sean precisamente los de umbrales de los nodos intermedios (Sepal-Length: 5.6, Sepal-Width: 3, **Petal-Length: 4.9**, **Petal-Width: 0.6**), la explicación resulta más extensa:

```
Iris is type Setosa because its
petal-width is low.
However, this iris may be also
Virginica because its petal-width
is quite close to the split value
(0.6).
It may be also Versicolor because
its petal-width and petal-length
are quite close to the split values
(0.6 and 4.9, respectively). For
these specific values it is just
as likely to be Virginica and
```

Generación Automática de Explicaciones en Lenguaje Natural para Árboles de Decisión de Clasificación

B. López-Trigo, Jose M. Alonso, A. Bugarín
Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS),
Universidade de Santiago de Compostela,

cases.

IV. Algunos Casos de Uso

Una vez descritos los elementos que componen cada explicación, veremos en esta sección dos ejemplos completos, con los que ilustraremos el funcionamiento de nuestra propuesta paso a paso. En ambos casos se aprenden clasificadores utilizando el algoritmo C4.5 [10], en la implementación disponible en Weka (J48) [16], [17]. Tanto los dos ejemplos mostrado (IRIS y FLAVIA), como otros disponibles, se pueden reproducir con el servicio Web ExpliClas [1] (Fig. 1).

IV-A. Conjunto de datos IRIS

El conjunto de datos IRIS (uno de los más conocidos del repositorio [18]) está formado por 150 instancias, 4 atributos numéricos y 3 clases. El árbol de clasificación generado por Weka (Fig. 2) está formado por 9 nodos totales, 5 de ellos nodos-hoja que deciden la clasificación y los 4 nodos restantes con las condiciones (comparaciones sobre los valores de los atributos) para decidir la clasificación. Se trata, por tanto, de un árbol simple que utilizaremos como primer ejemplo.

La explicación global generada en este caso es la siguiente:

```
    There are 3 types of iris:
Setosa, Virginica and Versicolor.
This classifier is very reliable
because correctly classified
instances represent 96%.
```

La explicación local, para la instancia de la Fig. 3 (Sepal-Length: 5.6, Sepal-Width: 3, Petal-Length: 4.1, Petal-Width: 1.3) es la siguiente:

```
    Iris is type Virginica because
its petal-length and petal-width
are medium.
```

En este caso, la explicación consiste en indicar los valores lingüísticos correspondientes a los valores numéricos de los atributos que han dado lugar a la clasificación, tal y como se detalla en la figura.

Sin embargo, si tomamos una instancia cuyos valores sean precisamente los de umbrales de los nodos intermedios (Sepal-Length: 5.6, Sepal-Width: 3, **Petal-Length: 4.9**, **Petal-Width: 0.6**), la explicación resulta más extensa:

```
    Iris is type Setosa because its
petal-width is low.
However, this iris may be also
Virginica because its petal-width
is quite close to the split value
(0.6).
It may be also Versicolor because
its petal-width and petal-length
are quite close to the split values
(0.6 and 4.9, respectively). For
these specific values it is just
as likely to be Virginica and
```


'Iris is type Setosa because its petal-width is low.
However, this iris may be also Virginica because its petal-width is quite close to the split value (0.6).
It may be also Versicolor because its petal-width and petal-length are quite close to the split values (0.6 and 4.9, respectively).'

External Demonstration

Pre-text	Given an instance with features: sepal length (cm) = 5.6, sepal width (cm) = 3.0, petal length (cm) = 4.9, and petal width (cm) = 0.6, a good explanation for why the instance was classified as virginica is:
Original	'By evaluating the feature values, it is possible to observe that both petal length and petal width are high. This means that by following the decision tree path, the instance should be classified as virginica, although the tree is not very confident in this result.'
New	'Iris is type Setosa because its petal-width is low. However, this iris may be also Virginica because its petal-width is quite close to the split value (0.6). It may be also Versicolor because its petal-width and petal-length are quite close to the split values (0.6 and 4.9, respectively).'

External Demonstration

Pre-text	`Given an instance with features: sepal length (cm) = 5.6, sepal width (cm) = 3.0, petal length (cm) = 4.9, and petal width (cm) = 0.6, a good explanation for why the instance was classified as virginica is:`
Original	`'By evaluating the feature values, it is possible to observe that both petal length and petal width are high. This means that by following the decision tree path, the instance should be classified as virginica, although the tree is not very confident in this result.'`
New*	`'Iris is type Setosa because its petal-width is low. However, this iris may be also Virginica because its petal-width is quite close to the split value (0.6). It may be also Versicolor because its petal-width and petal-length are quite close to the split values (0.6 and 4.9, respectively).'`

**We also changed the tree to be the same as in López-Trigo et al. case*

External Demonstration - Response

The decision tree concluded that the given example is virginica with a confidence of 97.06% because its petal width is high, which is a key factor in determining the class. Additionally, its petal length is high, which also contributes to the classification as virginica. These features align closely with the criteria used by the decision tree to classify instances as virginica.

External Demonstration - Response

The decision tree concluded that the given example is virginica with a confidence of 97.06% because its petal width is high, **which is a key factor in determining the class**. Additionally, its petal length is high, **which also contributes to the classification** as virginica. **These features align closely with the criteria used by the decision tree** to classify instances as virginica.

Original:

The decision tree concluded that the given example is virginica with a confidence of 97.06% because the petal length is high and the petal width is also high. This combination of high petal length and petal width led the decision tree to classify the instance as virginica with high confidence.

External Demonstration 2

Pre-text	Given an instance with features: sepal length (cm) = 5.6, sepal width (cm) = 3.0, petal length (cm) = 4.9, and petal width (cm) = 0.6, a good explanation for why the instance was classified as virginica is:
New	'Iris is type Setosa because its petal-width is low. However, this iris may be also Virginica because its petal-width is quite close to the split value (0.6). It may be also Versicolor because its petal-width and petal-length are quite close to the split values (0.6 and 4.9, respectively).'
Instruction Changes	For answering the following question, do not refer to the underlying mechanics of the decision tree in any way, and only refer to the features using natural language. All the relevant features must be mentioned in the answer, but features that were not used by the tree should be ignored. ~~**Moreover, do not use any technical jargon or numerical values in the response and prefer to user terms like 'high' and 'low'.**~~

External Demonstration 2 - Response

The decision tree concluded that the given example is virginica with a confidence of 97.06% because its petal width (1.8 cm) is greater than the split value (1.7 cm), and its petal length (6.3 cm) is also greater than the split value (4.9 cm). These features align with the criteria used by the decision tree to classify instances as virginica.

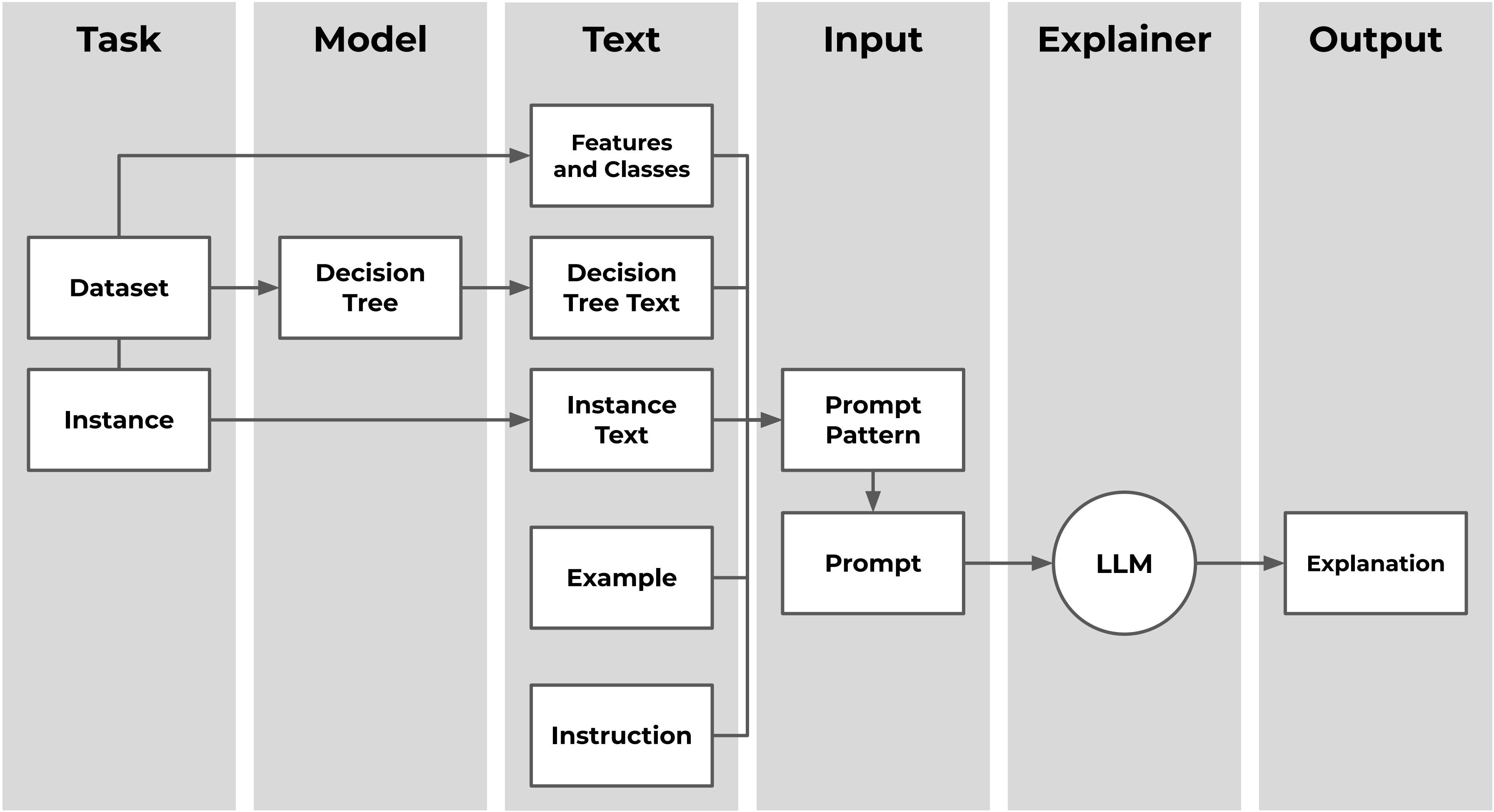
External Demonstration 2 - Response

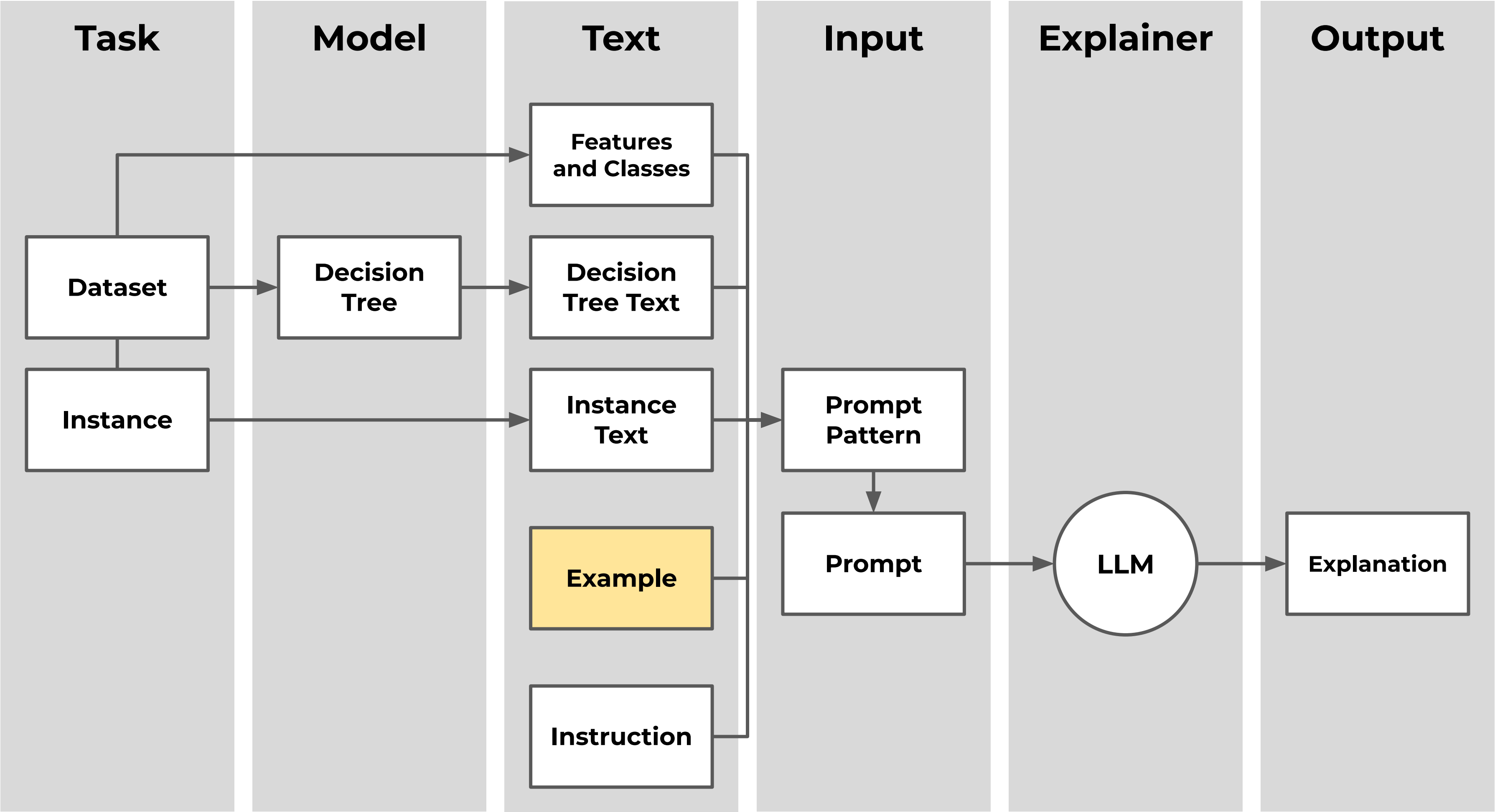
The decision tree concluded that the given example is virginica with a confidence of 97.06% because its petal width **(1.8 cm)** is greater than the **split value** **(1.7 cm)**, and its petal length **(6.3 cm)** is also greater than the **split value** **(4.9 cm)**. These features align with the criteria used by the decision tree to classify instances as virginica.

Demonstration:

'Iris is type Setosa because its petal-width is low. However, this iris may be also Virginica because its petal-width is quite close to the **split value** **(0.6)**. It may be also Versicolor because its petal-width and petal-length are quite close to the **split values** **(0.6 and 4.9, respectively)**.'

It seems that the explanation is very sensible to the instructions


Task
Model
Text
Input
Explainer
Output
Features and Classes
Dataset
Decision Tree
Decision Tree Text
Instance
Instance Text
Prompt Pattern
Example
Prompt
LLM
Explanation
Instruction



Task
Model
Text
Input
Explainer
Output
Features and Classes
Dataset
Decision Tree
Decision Tree Text
Instance
Instance Text
Prompt Pattern
Example
Prompt
LLM
Explanation
Instruction

Different Task Demonstration

New Task Tree - Wine	Demonstration
|--- color_intensity <= 3.82 | |--- ash <= 3.00 | | |--- od280/od315_of_diluted_wines <= 3.73 | | | |--- class: 1 | | |--- od280/od315_of_diluted_wines > 3.73 | | | |--- class: 0 | |--- ash > 3.00 | | |--- class: 0 |--- color_intensity > 3.82 | |--- flavanoids <= 1.58 | | |--- alcalinity_of_ash <= 17.65 | | | |--- class: 1 | | |--- alcalinity_of_ash > 17.65 | | | |--- class: 2 | |--- flavanoids > 1.58 | | |--- proline <= 724.50 | | | |--- class: 1 | | |--- proline > 724.50 | | | |--- class: 0	Given an instance of the **iris dataset** with features: sepal length (cm) = 3.6, sepal width (cm) = 0.8, petal length (cm) = 4.3, and petal width (cm) = 2.9, and a confidence value of 57.81 %, a good explanation for why the instance was classified as virginica is: 'By evaluating the feature values, it is possible to observe that both petal length and petal width are high. This means that by following the decision tree path, the instance should be classified as virginica, although the tree is not very confident in this result.'

Different Task Demonstration - Response

The decision tree classified the given example as class_0 with 100.00% confidence because the value of **color intensity is low,** and the **value of ash is high**. This led the tree to directly classify the example as class_0 without considering any other features.

Different Task Demonstration - Response

`The decision tree classified the given example as class_0 with 100.00% confidence because the value of` **`color intensity is low`**`, and the` **`value of ash is high`**`. This led the tree to directly classify the example as class_0 without considering any other features.`

The LLM generalizes the given demonstration to other tasks

Different Task Demonstration - Response

`The decision tree classified the given example as class_0 with 100.00% confidence because the value of` **`color intensity is low`**`, and the` **`value of ash is high`**`. This led the tree to directly classify the example as class_0 without considering any other features.`

The LLM generalizes the given demonstration to other tasks

Since it doesn't check for correctness, if the path is too complex, it might give incorrect explanations

Future Work

More tests with prompt engineering techniques

Apply the framework to other models

- Define a metric for explanation quality

Use this approach to surrogate Decision Trees methods (e.g. LORE)

Improve interaction with the user

Exploring Large Language Models Capabilities to Explain Decision Trees

Paulo Bruno De Sousa Serafim

Thank you!

paulobruno.github.io

paulo.desousa@gssi.it

GRAN SASSO SCIENCE INSTITUTE
SCHOOL OF ADVANCED STUDIES
Scuola Universitaria Superiore