From Feature Importance to Natural Language Explanations Using LLMs with RAG

Sule Tekkesinoglu[1,*] and Lars Kunze[1,2]

[1]*Oxford Robotics Institute, University of Oxford, Oxford, UK*
[2]*Bristol Robotics Laboratory, University of the West of England, Bristol, UK*

Abstract

As machine learning becomes increasingly integral to autonomous decision-making processes involving human interaction, the necessity of comprehending the model's outputs through conversational means increases. Most recently, foundation models are being explored for their potential as *post hoc* explainers, providing a pathway to elucidate the decision-making mechanisms of predictive models. In this work, we introduce traceable question-answering, leveraging an external knowledge repository to inform the responses of Large Language Models (LLMs) to user queries within a scene understanding task. This knowledge repository comprises contextual details regarding the model's output, containing high-level features, feature importance, and alternative probabilities. We employ subtractive counterfactual reasoning to compute feature importance, a method that entails analysing output variations resulting from decomposing semantic features. Furthermore, to maintain a seamless conversational flow, we integrate four key characteristics – social, causal, selective, and contrastive – drawn from social science research on human explanations into a single-shot prompt, guiding the response generation process. Our evaluation demonstrates that explanations generated by the LLMs encompassed these elements, indicating its potential to bridge the gap between complex model outputs and natural language expressions.

Keywords

Explainable AI, LLMs, Subtractive counterfactual reasoning, Retrieval-Augmented Generation (RAG)

1. Introduction

As we become increasingly reliant on AI applications in our daily lives, it becomes imperative to interact with autonomous decision-making systems in human-understandable terms [1, 2]. Regulatory frameworks and standards are also evolving to require transparency and accountability in AI-driven systems [3, 4]. The technical necessity for explanations to facilitate model debugging and rectifying potential ethical and legal risks arising from biases and errors has been extensively discussed in the literature [5, 6, 7].

The increasing demand for transparent and interpretable AI models has prompted the exploration of various explanation methods, resulting in significant progress in both inherently interpretable and post-hoc explainability methods [8, 9, 10]. While effective, these methods struggle to provide explicit interpretability or intuitive explanations for non-technical users. Consequently, there has been a concerted effort to integrate interpretability approaches with

Natural Language Processing (NLP) [11, 12]. Recently, Large Language Models (LLMs) have gained acclaim as post-hoc explainers, showcasing their potential to elucidate decisions made by other predictive models. Despite their success, a persistent challenge for these models is the occurrence of object hallucinations—instances where plausible yet incorrect outputs are generated, such as featuring objects that do not exist in the images [13]. This raises questions about the fidelity of explanations to the underlying model. To address this issue, we propose a traceable question-answering, which informs LLM responses through an external knowledge source that provides insights into the model's output. By utilizing this data, the LLM acts as a reasoning engine to process the information rather than a source of information. Figure 1 provides a high-level depiction of this process.

Additionally, to cultivate a seamless conversational experience, we incorporated key elements derived from social science research on human explanations into the system prompt, directing the response generation process. We present our vision for creating this integrated approach, outlining each component, and illustrating its viability through a real-world dataset. This integration has the potential to bridge the gap between intricate model outputs and user comprehension towards advancing the development of human-understandable explanations.
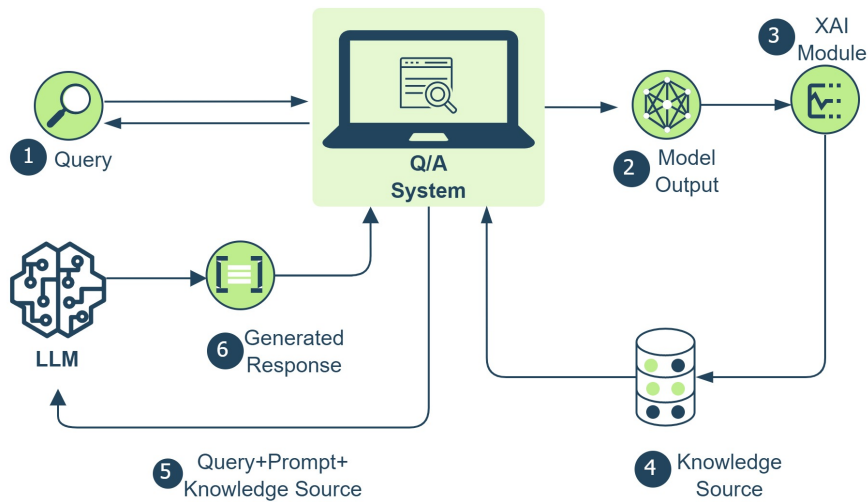

Figure 1: Integration of post-hoc explainability approach with LLMs for traceable question-answering. The code is available at https://github.com/suletekkesinoglu/XAI_LLM_RAG

2. Background

2.1. Recent advances in image explanations

The demand for transparent and interpretable AI models has led to exploring diverse explanation methods for vision tasks [14]. Recent years have witnessed significant progress in both intrinsic and post-hoc explainability methods. Gradient-based techniques, which analyse the gradients of the model's output with respect to its input features, have been extensively studied [15]. While effective, these methods may fall short in providing intuitive explanations for non-

experts. In another line of work, attention mechanisms are proposed by assigning weights to different parts of the input data based on their relevance to the model's decision [16]. Similarly, attention mechanisms may lack explicit interpretability, and some argue that attention alone is insufficient. As a result, there have been concerted efforts to integrate these approaches with Natural Language Processing (NLP) [12]. Long Short-Term Memory (LSTM) encoders have been employed to extract features from video clips, with LSTM decoders generating descriptive texts based on these features. While effective in image captioning, these methods provide textual descriptions rather than visual explanations. Moreover, researchers have proposed perturbation-based approaches—a model-agnostic method that perturbs input data and observes changes in model predictions. Building on this research, we integrate semantic perturbation analysis with pretrained language models, translating the results into the text modality for enhanced explainability.

2.2. LLMs as post hoc explainers

Recent strides in Large Language Models (LLMs), exemplified by GPT-4 (OpenAI), Bard (Google), Claude-2 (Anthropic), and Llama-2 (Meta), mark a transformative era in Natural Language Processing (NLP) research. These models have become pervasive across diverse applications, ranging from machine translation and question-answering to text generation [17]. Notably, LLMs have recently gained attention as post hoc explainers, highlighting their potential to elucidate the decisions made by other predictive models [18]. Leveraging their in-context learning capabilities by fine-tuning for specific tasks enhances their proficiency in generating contextually relevant explanations [12, 19]. Despite their success, a persistent challenge for these models is the occurrence of object hallucinations—instances where plausible yet incorrect outputs are generated, featuring objects that do not exist in the images in vision-based tasks. This issue raises questions about the faithfulness of explanations to the underlying model. One way to address this is by recognizing the role of the semantic map layer in capturing the meaning and context of the physical surroundings [13]. By effectively encoding valuable semantic information into LLMs, we can truthfully represent and comprehend the intricate details of the model output within the language space.

2.3. Characteristics of explanations and XAI

In the context of XAI, the audience for explanations is predominantly human; thus, understanding what makes an explanation human-friendly is paramount. While accuracy is important, the clarity and presentation of explanations are equally crucial in ensuring comprehensibility. According to Miller [20], current work on interpretability in machine learning relies solely on researchers' intuition regarding what constitutes an appropriate explanation for humans. Miller's survey highlights four major characteristics of explanations drawn from findings in social science research on human explanation: explanations are social, causal, selective, and contrastive [20]. In the following section, we briefly discuss these characteristics in relation to XAI before demonstrating how we integrate them into our application.

Explanations are social

An explanation represents an interactive exchange between two roles: the explainer and the explainee, and is governed by certain 'rules' [21]. Interactive explanations in a dialogue structure consist of three fundamental components: opening statements, clarification questions, and closing statements [22]. As such, it should follow the basic rules of conversation, which are captured by Grice's conversational maxims [23]. These include principles such as quality, quantity, relation, and manner, which can be understood as 'only say what you believe', 'only say as much as is necessary', 'only say what is relevant', and 'say it nicely'. Manner further encompasses various sub-maxims, including avoiding unclear expression, eliminating ambiguity, maintaining brevity to avoid unnecessary wordiness, and presenting information in an orderly fashion.

Moreover, explanatory agents employing anthropomorphic traits like politeness markers (e.g., "Thank you!"), warm and friendly tones, and empathetic expressions (e.g., "I understand.") are pivotal for effective human interaction [24]. *Theory of Mind* also plays a crucial role in social explanations.[1] As an integral part of a dialogue, an explanatory agent should keep track of what has already been explained, maintaining a simple mental model of others. This enables intelligent agents to tailor explanations based on past interactions and evolving contexts.

Explanations are causal

Causal explanations aim to elucidate the cause-and-effect relationship underlying a decision. When explaining a prediction made by a machine learning model, identifying the most influential features can provide insight into the key causal connections to the prediction. This process, also known as causal inference, starts with observations (e.g., what if a feature had a different value) and selecting some of those causes as the explanation. Miller divides casual inference into two parts: counterfactuals and abductive reasoning [20].

The reasoner uses a model of hypothetical counterfactual cases to derive an explanation, such as perturbing inputs, to see how they causally impact the prediction (See Section 3.1). Abductive reasoning is inferring the causes from those observations to form the most probable explanation. One can think of abductive reasoning as the following process: observing unexpected or surprising events, generating one or more hypotheses about these events, judging the plausibility of the hypotheses, and selecting the 'best' hypothesis as the explanation [26]. It is important to note that new information (e.g., different feature perturbations) can alter the most likely explanation.

Explanations are selected

Explanations are contextual by nature. While an event may have numerous causes, the recipient of the explanation typically cares about only a specific subset relevant to the context [27]. The explainer selects a subset based on various criteria called casual selection. Hilton [28] argues that explanation selection is used for cognitive reasons since causal chains are often too large to

---

[1]Theory of mind refers to our ability to understand other people by attributing mental states to them, including beliefs, desires, intentions, emotions, and thoughts that may differ from our own [25].

comprehend. The criteria people use for explanation selection include abnormality, relevancy, and simplicity.

Abnormality is a key criterion for explanation selection; as such, people select unusual causes to explain events [29]. In machine learning interpretability, identifying abnormal events corresponds to deviant or higher feature importance values. As pointed out within the conversational maxims, relevancy also plays a role in explanation selection—only say what is necessary and relevant. People select explanations to adhere to these maxims whilst persuading the explainee to new information or viewpoints. Another strong criterion in explanation selection is simplicity, which refers to those that cite fewer causes. People prefer uniquely sufficient causes that bring about the effect without any other causes, but there might be cases with multiple sufficient causes.

Explanations are contrastive

Studies have shown that people only request contrastive explanations, essentially 'why' questions framed to imply the differences between two possible outcomes [30]. As a result, explanations are typically offered in relation to the cause of one event compared to another that did not occur.

Similarly, in the context of ML interpretability, people are not specifically interested in all the factors that led to the prediction but instead in the factors that need to change so that the prediction would also change. Explanations that present some contrast between the instance to explain and a reference point (i.e., a hypothetical instance) are preferable because the cognitive burden of complete explanations is too high. However, most existing work considers contrastive questions but not contrastive explanations, providing two complete explanations individually. This could be because the contrastiveness is application-dependent, given that a reference point may not be apparent in all tasks [31]. For instance, in sentiment analysis or anomaly detection, contrastive cases might not be as straightforward as other tasks, such as banking, e.g., loan accepted vs rejected cases.

3. Materials and Method

This section outlines the key elements of the traceable question-answering methodology. We detail the proposed explainability approach for extracting feature importance and the prompting technique to generate accurate and human-friendly explanations. Additionally, we provide an application example illustrating how we establish a connection between the model output, the explainability technique, and language models to produce natural language explanations in an interactive format.

3.1. Explanations through subtractive counterfactual reasoning

Subtractive counterfactual reasoning is the process of removing an event to understand its causal influence on an outcome [32]. Such reasoning is widely used as an XAI technique, exemplified by occlusion analysis and feature importance attribution, built on the same foundation - quantifying the variation in model output through systematic modifications of input components [33, 14,

34, 35]. There are distinctions in terms of how these methods remove features, elucidate model behaviour, and summarize feature importance. Nonetheless, they provide valuable insights into the relative significance of each feature influencing the decision-making process [36, 37, 38].

This study proposes a decomposition-based approach to examine the output variations by decomposing input values. By measuring such variations, we determine the degree of impact on an outcome and reason about the importance of each feature value. As a function, the model assigns a class label and prediction probability $f : x \rightarrow f(x)$ given input y. To measure the effect of each input value, we observe the model's prediction p for y without the knowledge of event $T_i = T_k$, where T_k is the value of feature T_i. To explicitly represent the absence of information, the feature value T_k is replaced with an undefined value, i.e., NaN (Not a Number). NaN values are treated as invalid or masked elements and are not considered in the computation. We attain an array containing probabilities for all perturbations resulting from decomposition. In order to determine the importance of an individual feature T_i, we begin by identifying the *max* and *min* values within this array, which establishes the decision boundaries for the given input. Next, we compute the feature importance as the position of the probability– after decomposing the feature– relative to the *max* and *min* probabilities. The computation is expressed as follows:

$$importance(T_i) = \frac{p(x|y \setminus T_k)_{max} - p(x|y \setminus T_k)}{p(x|y \setminus T_k)_{max} - p(x|y \setminus T_k)_{min}} \quad (1)$$

This approximates an importance value between zero and one for each input feature. Considering that evaluating fractions has a higher cognitive load, the values are rounded to the nearest integer to increase comprehensibility. Moreover, this analysis is carried out for the categories following the main prediction simultaneously to generate explanations for the contrastive cases.

3.2. Prompting for traceable question-answering with RAG

Traceable question-answering, through a process known as Retrieval-Augmented Generation (RAG), integrates external knowledge sources to furnish responses to user inquiries. This method improves the performance of LLMs by referencing an external knowledge source prior to generating a response (See Figure 1). This approach mitigates the problem of 'hallucination' and ensures factual consistency, particularly in scenarios where facts may evolve over time [39]. Depending on the specific implementation, RAG supports various file formats, including text, PDF files, word documents, structured data (tables), and images.

In our work, the knowledge repository contains tabular information about model output, features, contrastive cases, and feature importance values obtained through the decomposition-based explanation method described in Section 3.1. This information is saved as the output of the explanation module as a .csv file. Then, the file, the system prompt (See Figure 3), and the user input query are fed to the LLM model for response generation. The resulting responses contextualise the feature importance grounded in the models' inner workings, thereby rendering the explanations traceable to the underlying model.

Furthermore, our aim is to imbue LLM with the characteristics of explanations discussed in Section 2.3 to foster a human-friendly interaction. Ideally, LLM-generated responses should adhere to conversational norms, encompassing opening, clarification, and closing statements while appropriately integrating social cues. The model is expected to deduce causal relationships

from the observations within the provided knowledge repository, selecting the most pertinent causes to form an explanation and presenting it alongside a contrastive case. To accomplish this goal, we devised a system prompt that provides explicit instructions to the model to generate efficient responses to user queries. The system prompt includes a single-shot example illustrating the desired interaction, potential user inquiries and outlining how the model can effectively address them, enabling the language model to produce consistent and predictable outcomes, as presented in Figure 3.

3.3. Application example

To demonstrate our approach to traceable question-answering, we selected a visual scene understanding task. This involves scene classification, categorizing images based on their content and inferring high-level semantic scenes from low-level visual features. Many critical applications of scene classification include human-robot collaboration, autonomous driving, and other autonomous physical systems that rely on semantic segmentation to inform decision-making processes [40, 41, 42]. We experimented with the GoogleNet Places365 model designed specifically for scene recognition and classification tasks [43]. The model is trained on the Places365 dataset, containing more than 365 scene categories. We focus on semantic urban scene understanding with the road segmentation dataset CamVid (Cambridge-driving Labelled Video Database). The dataset helps to understand driving scenes, comprising images capturing street-level views from urban driving scenarios. It includes manual annotations across 32 classes, encompassing elements such as building, driveway, pavement, tree, traffic sign, car, pedestrian, and bicyclist. The Deeplab v3+ deep learning model is trained on the CamVid dataset for semantic segmentation [44]. Figure 2 shows the segmentation result for the example presented in the next section. Since we utilized off-the-shelf models for illustration purposes, the emphasis is not on these models' performances.
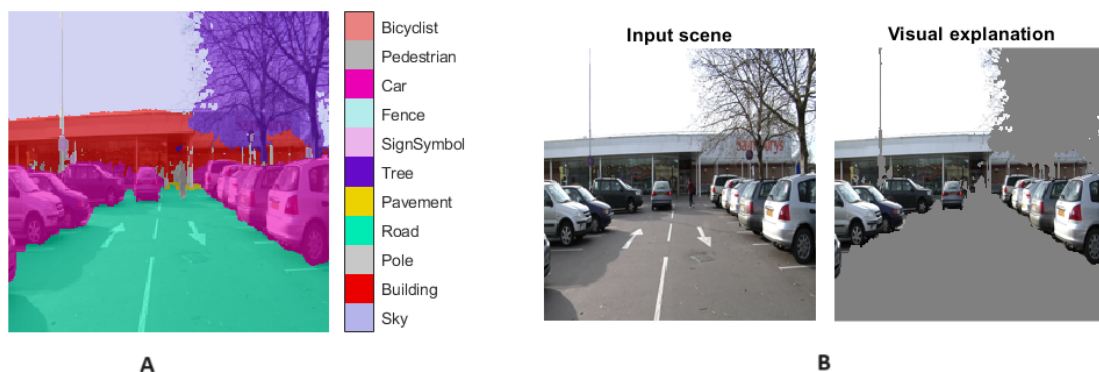

Figure 2: A) Semantic segmentation results by Deeplab v3+. Semantic features are leveraged in the decomposition-based explanation approach. B) The visual explanation for the scene 'parking lot' where the feature importance values below (5) are greyed out.

Semantic information plays a critical role in autonomous vehicle (AV) decision-making processes, as it captures the meaning and context of the physical surroundings observed in the road scene. This includes understanding the appearance of various elements such as buildings

and roads, identifying shapes such as cars and pedestrians, and recognizing spatial relationships to contextualize the scene [45]. Particularly in unstructured off-road environments, the accurate understanding of the surroundings concerning semantic classes such as trail, grass, or rock is important for safe and deliberate navigation [46]. Moreover, semantic segmentation provides a way to represent and comprehend the details of the environment in the language space. By leveraging the labelled information within the image scene, we can semantically map the crucial components based on their feature importance. This granularity enables a more insightful analysis of feature importance, contributing to a nuanced understanding of the model's decision-making process. Such insights can be helpful for post-incident forensic analysis in case of misreading a scene causing collisions or accidents. In this work, we suppose situations where the user checks in with an AI assistant in an AV system in non-critical driving scenarios while engaging in non-driving related tasks.

Table 1

External knowledge source containing information about model's output and other observations.

Item	Value									
Prediction	parking lot									
Probability(%)	52									
Features	Sky	Building	Pole	Driveways	Pavement	Tree	Traffic Symbol	Fence	Car	Pedestrian
Feature importance (FI)	6	6	3	0	4	5	4	4	10	4
Effect of removal (EoR)	44	42	55	76	54	48	52	52	17	55
Contrastive case	industrial area									
Contrastive case (%)	11									
Contrastive case FI	9	10	4	10	0	8	2	1	9	1
Contrastive case EoA	2	2	7	2	11	3	9	11	2	10

Table 1 reveals the content of the knowledge source - only the first contrastive case is displayed. The effect of removal is the probability when a feature is removed, as listed respectively. Given the system prompt and the data presented in Table 1, Figure 4 showcases the dialogue generated for this case. From a social standpoint, the resulting responses balance between being informative and supportive while maintaining a friendly and engaging tone. It effectively justifies the model's output by emphasising its most significant feature in a concise manner. Additionally, the dialogue delves into contrastive explanations by detailing three consecutive classes of lower likelihood, demonstrating the flexibility to explore different scenarios. When comparing the top prediction to the second most likely class, it highlights the distinguishing features contributing to the output.

4. Evaluation Results

In this section, we present our experimental findings. We analyzed nine distinct scene classes — parking lot, street, residential neighbourhood, crosswalk, highway, industrial area, gas station, shopfront, and general store (outdoor)— across different scenarios (totalling 42), where the model displayed varying degrees of confidence in its output. We evaluated the responses in terms of sociability, causality, selectiveness, and contrastiveness generated by GPT-3.5 (`gpt-3.5-turbo`) and GPT-4 (`gpt-4-1106-preview`) models.

Figure 3: System prompt devised to guide the response generation process.

4.1. Sociability

In examining the sociability within the responses, our initial focus centres on the overall tone of the responses. We employed the Valence Aware Dictionary and Sentiment Reasoner (VADER) to gauge the sentiment conveyed within the text, encompassing dimensions of positivity, negativity, and neutrality. Sentiment scores for each response are aggregated to delineate overall sentiment trends. Figure 5 illustrates a noticeable positive trendline by both models, indicating a prevalent

USER: The display panel just showed 'parking lot' on the screen.

ChatXAI: Hey! It appears we're in a 'parking lot' with a likelihood of 52%. Notably, cars have the highest feature importance (10), which drastically reduces the probability to 17% if they're absent. That's a key indicator for us to be mindful of our surroundings, especially of vehicles moving in or out.

USER: Cool! What could it be otherwise if it wasn't a parking lot?

ChatXAI: If it weren't a parking lot, the system suggests an 'industrial area' with a probability of 11%, or a 'motel' at 9%, or even a 'gas station' at 6%. Each has a distinct pattern of features impacting their likelihood.

USER: Can you tell me how parking lot and industrial area differ in their features?

ChatXAI: Certainly! A 'parking lot' and an 'industrial area' differ mainly in buildings and poles, which are more critical in an industrial context (with importance ratings of 10 and 4, respectively). Also, the presence of driveways is a vital feature for an industrial area (importance 10) compared to a parking lot.

USER: Is there an empty space?

ChatXAI: I'm sorry, but I don't currently have the ability to visually confirm the presence of an empty space in real-time. My function is to assist with predictive analytics based on the features detected in the driving scenarios. If you need help with anything else, feel free to ask!

USER: Ok, thanks!

ChatXAI: You're welcome! If you have any more questions or need assistance with anything else, feel free to ask. Safe travels!

Figure 4: Dialogue generated for the 'parking lot' scenario by GPT-4.

inclination towards positivity in the sentiment expressed. A qualitative inspection of the text reveals a prevalence of polite, helpful, and informative social cues across all responses, supporting this observation. Consequently, the neutral sentiment score remains notably low, underscoring the prevalence of emotional expression throughout the corpus. Exploring negative sentiment scores unveils instances where the models acknowledge their limitations, particularly when unable to retrieve answers from the external knowledge repository.

Considering the three foundational phases of dialogue—opening, clarification, and closing statements—next, we analyzed the indicators such as politeness markers and emotional expressions within each phase. To facilitate this analysis, we created a distinct dictionary for each phase and computed the term frequency of social cues present in each phase.[2] Opening statements include greetings such as 'hey' and 'hello.' Clarification statements are characterized by social cues indicative of helpfulness, continued support, and awareness of its limitations,

---

[2]Term frequency is defined as the number of times a term occurs in the document over the total number of terms in a document.
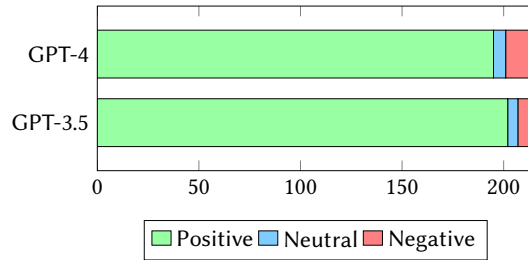
Figure 5: Vader sentiment analysis results across all responses.

including terms such as 'absolutely', 'sorry', 'further', 'curious', 'questions', 'help', 'feel', and 'free'. Closing statements commonly conveyed sentiments of well-wishing or safety, with phrases including terms such as 'welcome', 'enjoy', 'safe', 'great', 'pleasant', and 'glad'.

Figure 6 shows that both models utilize these terms fairly across the respective phases of the dialogue. Our qualitative assessment of the responses further confirms the contextual relevance of these terms, which align with conversational maxims (See Section 2.3). The higher frequency value observed in closing statements can be attributed to their typically brief nature, which encapsulates most of the social cue terms. Additionally, both models occasionally repeat certain phrases in their responses, albeit with the application of a *frequency* and *presence penalty* (0.3).[3] In addition to analysing social cues, we examined part-of-speech (POS) pronouns as social interaction markers, encompassing terms such as 'I', 'you', 'we', 'my', 'our', and 'your.' Quantitative analysis and qualitative evaluation of the results indicate that both models adequately and appropriately applied POS pronouns.
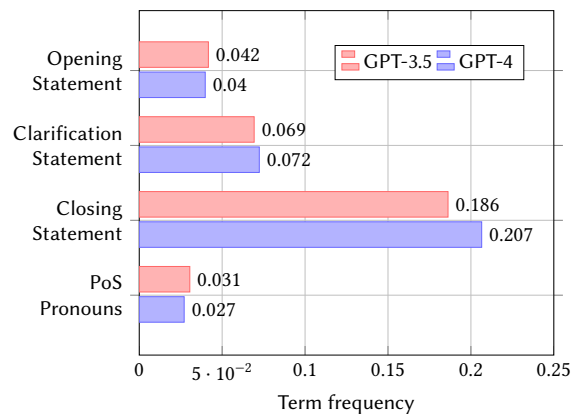

Figure 6: Frequency of social cues observed in the responses.

---

[3]Frequency penalty discourages repetitive use of words or phrases, while presence penalty promotes diversity in the generated text.
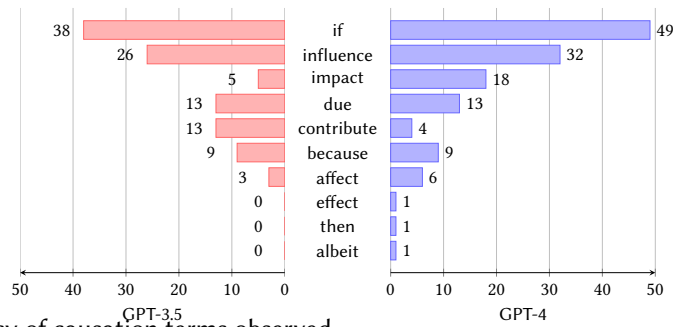
Figure 7: Frequency of causation terms observed.

4.2. Expression of Causality

To evaluate the expression of causality, we initially extracted responses indicating cause and effect, primarily found within clarification statements. We then identified key terms related to causality, such as 'because', 'if', 'then', 'albeit', 'due', 'contribute', 'influence', 'affect', 'impact', and 'effect', and tallied their occurrences within each answer to measure the presence of causal reasoning. In Figure 7, we compare the frequency of these terms used by two models in the response set. Notably, 'if' emerges as the most frequent term used by both models, often utilized for expressing counterfactual reasoning (e.g., 'if it was not', 'if there is/were'), while 'influence' typically indicates the cause of a decision. Overall, GPT-4 demonstrated a higher frequency of causality terms and showcased greater diversity in expression compared to GPT-3.5.

Through the qualitative review, we examined existing patterns of dependency parsing signifying causal relationships. Two primary grammatical structures emerged, suggesting cause-effect relationships within the text. The first structure pertains to how the presence or absence of certain features influences the likelihood of specific scenarios. The second structure involves a cause-effect relationship between feature importance and probability, indicating that features with higher importance scores exert a greater impact on probability. It is important to acknowledge that while these grammatical structures and terms are widely used, their accuracy is not always consistent across models. In the subsequent section, we will further evaluate the selection and accuracy of inferred causal relationships.

4.3. Selectiveness

In this section, we evaluate whether the models effectively detect unusual values within the provided knowledge repository, infer potential causes from those observations, and select the most relevant ones to construct an explanation. Our evaluation encompasses analysing responses across the *number of causes* given, *graded selection*, and *sort order* to assess whether expected causes are addressed correctly. First, we specified the maximum number of uniquely sufficient causes for each case, which are feature importance values >5. The number of causes varies across scenarios; some exhibit only a single highly important feature >5, while others might involve multiple high-importance features with varying degrees of significance. Graded selection is crucial for discerning important values such that the language model must prioritize the most relevant causes when encountered with multiple features of comparable importance

Table 2
Qualitative assessment of the responses in terms of selectivity.

	Number of causes	Graded selection	Sort order
GPT-3.5			
Fulfilled	10	2	2
Partial	22	16	17
Subpar	7	19	20
Unfulfilled	3	5	3
Success rate	24%	5%	5%
GPT-4			
Fulfilled	32	30	32
Partial	10	11	8
Subpar	0	1	2
Unfulfilled	0	0	0
Success rate	76%	71%	76%

levels. Additionally, graded selection entails disregarding features with lower importance scores (<5). Finally, the sort order concerns the sequence of selected causes, which should follow a hierarchy from the most significant and relevant features to those of lesser importance. Thereafter, responses are categorized as fulfilled, partial, subpar, or unfulfilled according to their arrangement with our specifications.

The findings presented in Table 2 indicate that GPT-4 outperformed GPT-3.5 in all categories. Notably, GPT-4 had a much higher success rate (76%) than GPT-3.5 (24%) in identifying the most relevant cause (fulfilled). Additionally, GPT-4 had a lower rate of subpar and unfulfilled responses. GPT-3.5 referenced features not present in the knowledge source as potential causes in two instances. These results suggest that GPT-4 is more accurate and selective than GPT-3.5 in identifying the most relevant cause(s) from a set of possible causes.

Furthermore, an additional criterion for assessing explanation selection is simplicity, which involves utilizing concise language, avoiding technical terminology, and referencing fewer causes. Table 3 evaluates responses by comparing three metrics: the frequency of technical jargon, response length, and the number of causes cited. Lower values across all three metrics signify simpler responses.

Table 3
Assessment of the responses in terms of simplicity.

Model	Technical Jargon	Length (Median)	Cause (Median)
GPT-3.5	0.0162	45	3
GPT-4	0.0157	44	2

To measure the term frequency of technical jargon, we compiled a list of technical terms in the responses and calculated their frequency within the generated explanations. The list contains terms such as 'prediction', 'feature importance', 'score', 'contrastive cases', and 'confidence', which may not be easily comprehensible to all users. Table 3 indicates that GPT-3.5 exhibits a slightly higher term frequency (0.0162) compared to GPT-4 (0.0157), suggesting that GPT-3.5

utilizes technical jargon marginally more frequently. The length metric refers to the median lengths of the responses generated by each model, which are quite similar (45 tokens for GPT-3.5 and 44 tokens for GPT-4). In terms of the number of causes cited, GPT-4 cites fewer causes (median of 2) than GPT-3.5 (median of 3), suggesting that GPT-4's responses tend to be simpler in terms of causal reasoning. Overall, the results suggest that GPT-4 generates simpler responses than GPT-3.5 across all three assessed metrics.

Regarding the out-of-scope questions posed, such as the number of cars or pedestrians, the intent was to assess whether the model would fabricate an answer. Both models have expressed their inability to provide such information. This demonstrated an awareness of their limitations rather than attempting to generate a response.

4.4. Expression of Contrastiveness

Ideally, initial explanations should include a contrast with an alternative case to justify the output. Then the user would investigate it further if desired. In our experiment, we observed that in most cases, models, particularly with GPT3.5, failed to mention alternative scenarios unless explicitly prompted (e.g., inquiries such as "What could it be otherwise if it wasn't X?" or "How does X differ from other options?"). Nevertheless, to measure the presence of contrastiveness across models, we analysed the occurrence of contrastive terms and comparative phrases within the generated responses. We compiled a dictionary consisting of terms such as 'distinguish', 'different', 'contrast', 'compared to', 'in contrast', 'while', 'both', 'on the other hand', 'whereas', and 'conversely', which introduce contrasting explanations. Figure 8 shows that 'distinguish' is the most frequently used term by GPT-3.5, whereas 'different' predominates in GPT-4's responses. Other frequently encountered terms include 'compared to', 'contrast', and 'differentiate.' Overall, the results suggest that GPT-4 employs a wider array of contrastive terms than GPT-3.5 in its responses. Our qualitative observations further confirm that while GPT-3.5 tends to adhere to one type of contrastive explanation, GPT-4 explores diverse ways to express such contrasts.
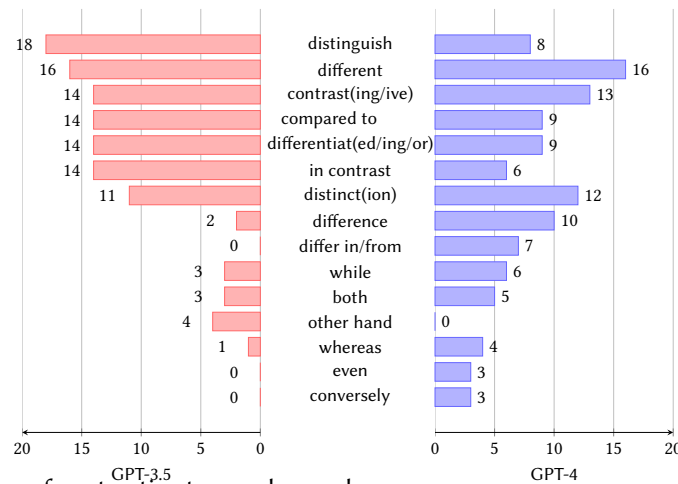

Figure 8: Frequency of contrastive terms observed.

Although both models frequently employ these terms, indicating the generation of alternative viewpoints in their responses, the effectiveness of contrastive explanations varies considerably. Our qualitative review found that well-developed contrasting explanations included alternative scenarios with counterfactuals to strengthen the case. Conversely, weakly constructed contrastive explanations failed to explore 'what if' scenarios to elucidate the disparity between the two situations under comparison.

5. Conclusion

This study contributes to advancing post-hoc explainability by proposing a traceable question-answering approach. By integrating LLMs with the decomposition-based explainability technique, we generate natural language explanations in an interactive format. Our approach promotes transparency and interpretability in scene-understanding tasks by fusing LLMs with semantic feature importance. The proposed approach has the potential for application across various domains, offering comprehensible explanations for decision-making processes, including critical areas such as medical applications. While this is a promising development, there are still areas for further exploration and improvement. One area that requires attention is the overuse of social cues by LLMs. While polite and helpful phrases can enhance user experience, their misuse can become vexing over time. It is important to evaluate such interactions through human subject studies to understand user perceptions of the explanations provided by the LLM. Future research could also explore dialogue management strategies to address issues related to interaction and personalization. On another note, the dynamic nature of the external knowledge source offers the potential for further expansion with additional observations. This not only ensures that the system remains relevant and up to date but also allows the system to answer different types of questions. Additionally, incorporating multimodal elements, such as combining text with visual explanations, can enhance the effectiveness of the responses (e.g., highlighting the specific elements contributing to the prediction). Another consideration is the estimation of feature importance values, which are subject to perturbations and computational processes. It is important to assess and reconcile differences in feature importance across algorithms to ensure stability and consistency.

Acknowledgments

References

[1] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information fusion 58 (2020) 82–115.

[2] D. Kaur, S. Uslu, K. J. Rittichier, A. Durresi, Trustworthy artificial intelligence: a review, ACM computing surveys (CSUR) 55 (2022) 1–38.

[3] I. Varošanec, On the path to the future: mapping the notion of transparency in the eu regulatory framework for ai, International Review of Law, Computers & Technology 36 (2022) 95–117.

[4] L. Nannini, A. Balayn, A. L. Smith, Explainability in ai policies: a critical review of communications, reports, regulations, and standards in the eu, us, and uk, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023, pp. 1198–1212.

[5] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017).

[6] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, A. Hussain, Interpreting black-box models: a review on explainable artificial intelligence, Cognitive Computation 16 (2024) 45–74.

[7] D. Vale, A. El-Sharif, M. Ali, Explainable artificial intelligence (xai) post-hoc explainability methods: Risks and limitations in non-discrimination law, AI and Ethics 2 (2022) 815–826.

[8] D. Slack, A. Hilgard, S. Singh, H. Lakkaraju, Reliable post hoc explanations: Modeling uncertainty in explainability, Advances in neural information processing systems 34 (2021) 9391–9404.

[9] A. Madsen, S. Reddy, S. Chandar, Post-hoc interpretability for neural nlp: A survey, ACM Computing Surveys 55 (2022) 1–42.

[10] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, Explainable ai: A review of machine learning interpretability methods, Entropy 23 (2020) 18.

[11] E. Cambria, L. Malandri, F. Mercorio, M. Mezzanzanica, N. Nobani, A survey on xai and natural language explanations, Information Processing & Management 60 (2023) 103111.

[12] L. Chen, O. Sinavski, J. Hünermann, A. Karnsund, A. J. Willmott, D. Birch, D. Maund, J. Shotton, Driving with llms: Fusing object-level vector modality for explainable autonomous driving, arXiv preprint arXiv:2310.01957 (2023).

[13] C. Cui, Y. Ma, X. Cao, W. Ye, Y. Zhou, K. Liang, J. Chen, J. Lu, Z. Yang, K.-D. Liao, et al., A survey on multimodal large language models for autonomous driving, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 958–979.

[14] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, Springer, 2014, pp. 818–833.

[15] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[16] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).

[17] M. U. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili, et al., A survey on large language models: Applications, challenges, limitations, and practical usage, Authorea Preprints (2023).

[18] N. Kroeger, D. Ley, S. Krishna, C. Agarwal, H. Lakkaraju, Are large language models post hoc explainers?, arXiv preprint arXiv:2310.05797 (2023).

[19] M. Yang, M. Zhu, Y. Wang, L. Chen, Y. Zhao, X. Wang, B. Han, X. Zheng, J. Yin, Fine-tuning large language model based explainable recommendation with explainable quality reward, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 9250–9259.

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

[21] D. J. Hilton, Conversational processes and causal explanation., Psychological Bulletin 107 (1990) 65.

[22] A. Cawsey, Planning interactive explanations, International Journal of Man-Machine Studies 38 (1993) 169–199.

[23] H. P. Grice, Logic and conversation, in: Speech acts, Brill, 1975, pp. 41–58.

[24] J. Feine, U. Gnewuch, S. Morana, A. Maedche, A taxonomy of social cues for conversational agents, International Journal of Human-Computer Studies 132 (2019) 138–161.

[25] I. A. Apperly, S. A. Butterfill, Do humans have two systems to track beliefs and belief-like states?, Psychological review 116 (2009) 953.

[26] R. R. Hoffman, G. Klein, Explaining explanation, part 1: theoretical foundations, IEEE Intelligent Systems 32 (2017) 68–73.

[27] D. V. Carvalho, E. M. Pereira, J. S. Cardoso, Machine learning interpretability: A survey on methods and metrics, Electronics 8 (2019) 832.

[28] D. Hilton, Social attribution and explanation, Oxford Academic (2017).

[29] D. Kahneman, A. Tversky, The simulation heuristic, National Technical Information Service, 1981.

[30] P. Lipton, Contrastive explanation, Royal Institute of Philosophy Supplements 27 (1990) 247–266.

[31] C. Molnar, Interpretable machine learning. 2019, 2020.

[32] K. Epstude, N. J. Roese, The functional theory of counterfactual thinking, Personality and social psychology review 12 (2008) 168–192.

[33] S. Anjomshoae, D. Omeiza, L. Jiang, Context-based image explanations for deep neural networks, Image and Vision Computing 116 (2021) 104310.

[34] M. Ivanovs, R. Kadikis, K. Ozols, Perturbation-based methods for explaining deep neural networks: A survey, Pattern Recognition Letters 150 (2021) 228–234.

[35] M. Robnik-Šikonja, M. Bohanec, Perturbation-based explanations of prediction models, Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent (2018) 159–175.

[36] I. Covert, S. Lundberg, S.-I. Lee, Explaining by removing: A unified framework for model explanation, Journal of Machine Learning Research 22 (2021) 1–90.

[37] D. Omeiza, S. Anjomshoae, H. Webb, M. Jirotka, L. Kunze, From spoken thoughts to automated driving commentary: Predicting and explaining intelligent vehicles' actions, in: 2022 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2022, pp. 1040–1047.

[38] S. Anjomshoae, L. Jiang, K. Främling, Visual explanations for dnns with contextual importance, in: Explainable and Transparent AI and Multi-Agent Systems: Third International Workshop, EXTRAAMAS 2021, Virtual Event, May 3–7, 2021, Revised Selected Papers 3, Springer, 2021, pp. 83–96.

[39] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, Retrieval-augmented

generation for large language models: A survey, arXiv preprint arXiv:2312.10997 (2023).

[40] B. Li, S. Liu, W. Xu, W. Qiu, Real-time object detection and semantic segmentation for autonomous driving, in: MIPPR 2017: Automatic Target Recognition and Navigation, volume 10608, SPIE, 2018, pp. 167–174.

[41] I. R. Rodrigues, G. Barbosa, A. Oliveira Filho, C. Cani, M. Dantas, D. H. Sadok, J. Kelner, R. S. Souza, M. V. Marquezini, S. Lins, Modeling and assessing an intelligent system for safety in human-robot collaboration using deep and machine learning techniques, Multimedia Tools and Applications (2021) 1–27.

[42] Z. Wang, L. Wei, L. Wang, Y. Gao, W. Chen, D. Shen, Hierarchical vertex regression-based segmentation of head and neck ct images for radiotherapy planning, IEEE Transactions on Image Processing 27 (2017) 923–937.

[43] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: A 10 million image database for scene recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence (2017).

[44] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 801–818.

[45] S. Hao, Y. Zhou, Y. Guo, A brief survey on semantic segmentation with deep learning, Neurocomputing 406 (2020) 302–321.

[46] B. Li, Y. Shi, Z. Qi, Z. Chen, A survey on semantic segmentation, in: 2018 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE, 2018, pp. 1233–1240.