Delft University of Technology

Master Thesis

Contrastive Self-Explanation Method (CoSEM): Generating Large Language Model Contrastive Self-Explanations

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science *in* Computer Science

Author:
Radek Kargul

Supervisors:
Jie Yang, Sarah E. Carter, Stefan Buijsman

Thesis Committee:

Chair:	Sole Pera
Committee Member:	Myrthe Tielman
Committee Member:	Jie Yang
Committee Member:	Sarah E. Carter
External Committee Member:	Stefan Buijsman

September 2024

Web Information Systems
Computer Science and Mathematics
Electrical Engineering, Mathematics and Computer Science (EEMCS), Delft University of Technology
Delft, the Netherlands

Contents

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my thesis supervisors, Jie Yang, Sarah E. Carter and Stefan Buijsman. Your guidance, encouragement, and insightful feedback have been invaluable throughout the entire process of this research. Your constructive critiques and thoughtful suggestions have contributed significantly to the quality of this work. Your expertise and perspectives have broadened my understanding of the subject, and I am profoundly grateful. I would also like to thank Sarah for helping me with all the logistics required to complete the user study (part of my thesis), the inspiring discussions, and simply being there when I needed encouragement. A special thank you goes to Jie for providing me with the resources and support necessary to complete my research. I also sincerely thank Sole Pera for chairing my thesis defence and Myrthe Tielman for being a part of the thesis committee members.

I would also like to thank my friend Nioosha for being there to share ideas, provide feedback, and offer support during challenging times. I would also like to thank all my friends, especially Clàudia, Nioosha, Niko and Aaron, who attended fitness classes with me and supported my journey to becoming an instructor. The camaraderie and shared experiences have made this journey all the more rewarding.

Lastly, I am deeply indebted to my parents and sister for their unwavering support and patience throughout my academic journey. Thank you for your constant encouragement and always believing in me.

Thank you all for making this journey possible and deeply meaningful.

Abstract

Large language models (LLMs) are widely used tools that assist us by answering various questions. Humans implicitly use contrast as a natural way to think about and seek explanations (i.e., "Why A and not B?"). Explainability is a challenging aspect of LLMs, as we do not truly understand how good the LLM answers are. The challenge is understanding to what extent LLMs can generate effective contrastive self-explanations for users. We introduce the Contrastive Self-Explanation Method (CoSEM) to narrow the gap between LLMs and explainability. It generates contrastive self-explanations and evaluates them through automation and a user study on generality, usefulness, readability, and relevance. Our results indicate that LLMs are capable of generating effective contrastive self-explanations. Lexical analysis of contrastive explanation indicates that explanations are not less general than the text those explain, and semantic analysis shows that more complex models generalize self-explanations more consistently. Although it is challenging to evaluate contrast in self-explanations semantically, user study shows that some models (Llama3-8B) help understand the contrast. Moreover, task selection affects how readable users find the explanations, where more self-explanations on general topics (movie reviews) are more readable than more specific topics (medical diagnoses). Lastly, some models, such as Llama3-8B, excel at generating contrastive self-explanations that contain relevant information regarding input text.

1 Introduction

Explainability is a challenging aspect of natural language processing (NLP) in large language models (LLMs). LLMs are complex "black-box" systems with opaque inner workings, making model interpretation difficult. GPT-4 is a pre-trained language model for generating text from given inputs. It is a closed-source model, and its internals are not publicly available. This multimodal model has human-level performance on specific complex professional and academic benchmarks [1]. Such models can classify sentiment presented in text (e.g., opinion mining), categorize text into predefined classes (e.g., spam detection), answer natural language questions (e.g., intelligent assistants), and generate human-like text [6]. An instruction-tuned version of GPT, ChatGPT, is trained on instructions that make the model capable of conversing in a dialogue-like setup. Such models are trained on vast amounts of data, giving them broad knowledge to leverage to generate responses to user prompts. However, this comes with a tradeoff. As LLMs become larger and more capable, they become more complex and less interpretable to humans. The sheer number of parameters in models like GPT-4 (over 1 trillion) makes it difficult to understand their inner workings and decision-making processes [5]. LLMs will only become more complex and advanced; therefore, we must try to understand them and provide better explanations for their decisions.

Non-contrastive question: Why did you steal from banks?

Humans implicitly assume a contrast: Because that's where **the money** is.

Contrastive question: Why did you steal from **banks** rather than **other places**?

Figure 1: Example of a human, a famous bank robber, assuming a specific contrast when being asked a non-contrastive question. The robber was asked a non-contrastive question, which he answered by having a specific contrast in his mind. The question he answered was contrastive.

Contrastive Explanations Humans implicitly use contrast as a natural way to think about and seek explanations (i.e., "Why A and not B?") [41]. Figure 1 shows an example of a famous bank robber answering a question in a contrastive manner with an explicit contrast in mind. Even though he was not asked, "Why do you rob banks rather than work?" he assumed a specific contrast that one could address with "Why do you steal from banks and not elsewhere?". [17] **Contrastive explanations** (CEs) involve comparing two concepts, ideas, or situations to highlight their differences - namely, *the fact* (A) and *the foil* (B) [34]. When asked, "Why are you home?" we can reply, "Because that is where I live." However, we cannot use the same answer if we add contrast (the foil), "Why are you home and not at work?". "Because I am sick and do not want to spread germs" would be more appropriate and accessible for us to understand, as the first answer might not be relevant if our boss asks about our whereabouts. Research has shown that CEs can help learners better understand new information when compared with non-CEs [63]. Much research focuses on generating CEs [8, 9, 38, 48, 60]. However, these approaches focus on methods that require internal access to the model, which is impossible when dealing with black-box LLMs.

Motivation The challenge we must address is understanding to what extent LLMs can generate effective contrastive self-explanations. The arduous part lies in understanding whether we, as users of such black-box systems, can use LLMs to explain concepts in a human-friendly (contrastive) manner in various domains (tasks). We should evaluate the effectiveness of contrastive self-explanations to make progress in making LLMs better companions for our daily life tasks. Apart from the models' faithfulness [23, 39] and explanations' minimality [9, 48, 60], we should understand whether these complex models can give contrastive self-explanations that address:

- *generality*, to keep the information on a general level without diving into deep specifics;
- *relevance*, to not deviate from the input's topic and prevent the introduction of unrelated information to mislead the explainee (e.g., hallucinations, misinformation);
- *usefulness*, to leverage the information and help understand the contrast in an explanation;
- *readability*, to ensure that users can understand and comprehend the information provided to them and use it to their advantage in learning.

Research Questions This work marks a significant step forward in the interpretability of complex AI systems, providing insights into their decision-making processes in a user-centric manner. Our research aims to address the following question **"To what extent can large language models generate effective contrastive self-explanations?"**, divided into the following sub-questions to evaluate the *effectiveness*:

- How does model complexity shape its *self-consistency* and *accuracy*?
- How well do contrastive self-explanations *generalize*?
- How *readable* are contrastive self-explanations?
- To what extent does task selection affect contrastive self-explanation?
- How useful (contrastive), readable, and relevant are deemed contrastive self-explanations to users?

Contribution We introduce the Contrastive Self-Explanation Method (CoSEM), which generates contrastive self-explanations and evaluates them on generality, usefulness, readability, and relevance. It is a black-box-model-friendly approach that does not require access to LLMs' internals. In our approach, to generate CEs, we do not use gradient-based perturbation but opt-in for model prompting instead [25, 63]. Very recent work on generating contrastive explanations for LLMs [36] also focuses on explaining black-box models through contrastive explanations. However, the authors of that work demonstrate how slightly modifying a prompt could lead to a different, potentially less desirable response from an LLM. In our work, we utilize self-explanations rather than explanations. We let various instruction-tuned models (Llama2, Llama3 [56], and Mistral [27]) generate contrastive self-explanations for each task. We then evaluate those explanations using an automated approach and a questionnaire (user study). These evaluations help us assess the challenges we are addressing in this study. To evaluate the models for *generality*, we have introduced a rule-based syntactic modification approach to changing text words while keeping their overall semantic meaning. This approach allows us to observe whether models can keep the information on a general level without going into specifics despite a syntactically modified input. We address *readability* through examining the level required to read the explanations, by counting the words per text, words per sentence, and syllables per word. Finally, we deploy a user study to address the *usefulness*, *relevance*, and *readability* of the self-explanations, which necessitates a human-centered study, as relying solely on automated methods without human input would be inadequate and could fail to capture the nuances of human interpretation and understanding.

As part of this work, we have proposed the following contributions:

- A novel non-gradient-based method through strategic model inference for generating contrastive self-explanations and for evaluating them automatically and through user feedback. The technique generates self-explanations without needing internal model access and is suitable for black-box environments.
- We have adapted an existing self-consistency checks method to CoSEM for contrastive self-explanations.

- We created a rule-based syntactic modification that changes a text paragraph and evaluates contrastive self-explanations for explanation generality.
- We proposed an inference method for generating contrastive and non-contrastive self-explanations through instruction-tuned LLMs.
- We covered self-explanations effectiveness across one general and one specific task, ensuring that the generated self-explanations are technically feasible, practically useful, and understandable to users.

Paper Structure This paper is organized into seven main sections following this introduction. Section 2 reviews existing literature and delineates three sub-categories: contrastive explanations, self-explanations, and the transition from white-box to black-box contrastive explanations, concluding with an overview of evaluation criteria. In Section 3, we discuss our approach to task selection using two datasets, evaluate model accuracy, verify self-consistency in contrastive explanations, generate modified explanations, and assess the efficacy of model self-explanations through various evaluations. Sections 4 and 5 present our results and discuss their implications, respectively. Section 6 addresses the limitations of our study and outlines potential directions for future research. The paper concludes with Section 7, summarizing our findings.

2 Related Work

2.1 Contrastive Explanations

Although counterfactual and contrastive explanations are closely related terms in XAI, they have various interpretations in the literature. Contrastive explanations, as acknowledged in [34, 42], are commonly structured as stating why a particular outcome (y) is reached for a given situation (x) rather than a different outcome (y'), by highlighting the differences in the values of specific features ($x_1, ..., x_n$) between the two scenarios. On the other hand, counterfactual explanations give an instance x, which has minimally changed to reach y'. Guidotti [19] claims that, in XAI, there is no difference between contrastive and counterfactual explanations, as in both cases, the aim is to find what would have changed the decision, either altering x or by comparing x with another instance. Both explanations look for minimal changes, even though the contrastive explanations look for a more constrained change, to the input for the decision of the black-box model to flip [14].

Counterfactual explanation: If your annual income had been €50,000 instead of €40,000, your loan application would have been approved.

Contrastive explanation: Your loan application was denied because your annual income is €40,000, while applicants with incomes of €50,000 or higher are approved.

Figure 2: Example of the distinction between a counterfactual and a contrastive explanation. A counterfactual explanation gives an instance of a problem that would change the contrast (from fact to the other hand, a contrastive explanation provides information on the features that need to change in order to achieve the change).

In this research, a distinction between contrastive and counterfactual explanations is made. To illustrate, consider a loan application scenario where a customer is denied a loan (Figure 2). A counterfactual explanation provides a specific change to the input (income) to achieve a different outcome (loan approval). In contrast, a contrastive explanation would focus on the difference between the approved and denied scenarios. This highlights the specific feature (income) that differentiates the outcome. Counterfactual examples are alternative scenarios that modify input data to achieve a different outcome from a model. They present "what-if" situations to help users understand and act on algorithmic decisions. Conversely, contrastive explanations seek to understand the differences between these examples rather than just generate them. The explanations focus on noticing distinctive differences between fact and foil cases.

2.2 Self-explanations

LLMs like ChatGPT can generate self-explanations. This enables them to articulate the reasoning behind their decisions, particularly in tasks involving text understanding and interpretation, such as sentiment analysis. However, the quality and faithfulness of these self-explanations can vary. Studies have found that while LLM self-explanations perform comparably to traditional explanation methods like occlusion and LIME, they can differ significantly regarding agreement metrics, suggesting they may not always entirely reflect the model's actual decision-making process [23, 39]. For example, in a sentiment analysis task, the LLM may explain its positive or negative sentiment classification by pointing to specific sentiment-laden words in the text [23].

Huang et al. [23] investigate how good the automatically generated self-explanations from LLMs are. Their work is one of the first to study LLM-generated feature attribution explanations. They obtain self-explanations in two ways: (1) generating an explanation before prediction and (2) generating the prediction first and then explaining it. They also evaluate their faithfulness based on metrics: comprehensiveness, sufficiency, decision flip rate under the removal of most important tokens, minimum fractions of tokens needing to be removed to cause decision flip, and rank correlation with model prediction after word deletion. The research shows that these explanations are comparable to traditional methods such as occlusion (the influence of removing a single word) and LIME (removing a random subset of words and applying linear regression to distribute scores to words).

However, they also notice that self-explanations are much cheaper to generate than explanations. Therefore, using self-explanations to explore usefulness seems to be a practical approach.

Madsen et al. [39] propose employing self-consistency checks to measure self-explanation faithfulness and see whether it reflects the model's behavior. For example, if LLM indicates certain words are essential for a prediction, it should not be able to make the exact prediction without these words. For the first time, they apply self-consistency to the counterfactual, importance measure, and redaction explanations, as those were only successfully applied to evaluate faithfulness. They find that faithfulness depends on the task, model, and type of self-explanation. For example, counterfactual explanations were more faithful for Llama2. In their methodology, the authors use three prompting setups: "Objective" (What is the...), "you" (What would you classify...), and "human" (What would a human classify...). However, their results indicate no significant difference between the settings. Moreover, in their perturbation method (when removing words predicted by a model), the authors replaced those spaces in input text with None, "[REDACTED]" or "[REMOVED]" tokens. However, there was no significant advantage to any of the settings. The authors conclude that LLMs' self-explanations should not be generally trusted, as their faithfulness can vary significantly. Similarly to their work, we do not consider closed models due to possible liability issues. Moreover, we avoid using SOTA models such as ChatGPT-4 and choose Llama3, Mistral, and Falcon models, which are all publicly available yet still offer good performance.

2.3 Contrastive Explanations for Black-Box LLMs Using Text Data

As LLMs become increasingly prevalent, the need to understand their reasoning grows clearer. Consequently, there is a vital need for additional research to improve explainability methods for text data. A well-explored path is through token importance. We distinguish between two such methods: gradient-based and perturbation-based token importance. In general, feature importance methods explain how relevant model features are concerning the model's prediction. These methods provide insight into which features the model is using to make its predictions. A great deal of such methods exist for tabular and image data types. Scott and Su-In [35] proposed the SHAP (Shapley Additive exPlanations) framework for tabular data. It assigns importance values to each feature for a specific prediction, aiming to address the challenge of interpreting highly accurate but difficult-to-interpret models. For image data, saliency maps highlight the regions of an image that are most salient or important for a particular task (e.g., classification, object recognition). Dhurandhar et al. [14] introduced the Contrastive Explanation Method (CEM), which generates instance-based local black-box explanations for classification models. The method aims to provide clear explanations by identifying features that should be minimally and sufficiently present or absent from the instance to be explained. This makes it suitable for images and tabular data without categorical features.

In a similar work, Luss et al. [36] developed the first contrastive explanation methods for LLMs operating only with black-box or query access. Their methods demonstrate how slightly modifying a prompt could lead to a different, potentially less desirable response from an LLM. They introduce two algorithms—one myopic and one budgeted—to effectively generate these explanations while managing computational resources. One main difference in their approach is the dependence on the infilling methods, which affect the quality of the contrast prompts that can be generated. We opt in for prompting a model with a contrastive question (i.e., Why *fact* and not *foil*?) to let the model produce a contrastive self-explanation rather than an explanation.

In a recent work by Tekkesinoglu et al. [55], the authors explore the generation of contrastive explanations as a part of their broader framework for providing natural language explanations using LLMs. They emphasize the importance of contrastive explanations in making an ML model's output more interpretable to users by explaining why a model chose one outcome over another. The approach enhances the transparency and trustworthiness of the explanations, making them more aligned with human reasoning processes. They utilize contrastive explanations to improve the interpretability of LLM-generated outputs while we evaluate generated model outputs.

Recently, there has been more focus on text data with the emergence of LLMs. Jacovi et al. [25] propose a method to produce contrastive explanation via a projection of the input representation, only to keep features that differentiate two potential decisions. By narrowing down the range of causal factors, such explanations simplify communication and alleviate the cognitive burden for both the one and the one being explained. They show that non-contrastive explanations address many tokens in the input space, while contrastive explanations focus only on a fraction of the tokens. They propose that non-contrastive explanations may not align with human expectations of the explanations' content, thereby complicating their interpretability by humans. The explanations can answer for which label and against which alternative label the feature is useful. Contrastive explanations are produced through low-level input token attribution (i.e., textual highlights) and high-level abstract content attribution (e.g., gender). Interventions select the causal factors from discrete or abstract features (in the input) for model explanations. Then, contrastive attribution selects factors that yield contrastive explanations. Those explanations are in the form of a dense representation of the input in the latent space. This representation is given by a projection operation that only keeps components that distinguish the facts from the foil. This work's main limitation is the lack of variations of interventions, comparing the fact to a set of foils. Moreover, the work focuses on highlighting tokens that contribute to answering a (contrastive) question. Similarly, the work of Yin and Neubig [63] focuses on model contrastive explanations using gradient-based attribution (input saliency). They measure the saliency of the model input based on the output to see what parts of the input affect the prediction and to what degree. They demonstrate that contrastive explanations significantly improve contrastive model simulatability for human observers.

Table 1: Summary of Methods for Explainability

Method	**Description**	**Key Points**
SHAP [35]	Shapley Additive exPlanations for tabular data	Assigns importance values to features for specific predictions, interpreting complex models
CEM [14]	Highlight regions of an image most salient for a task	Identifies features minimally and sufficiently present or absent for explanations, used for classification and object recognition in image data
Contrastive Explanations for LLMs [36]	Modifies prompts to generate contrastive explanations	Uses myopic and budgeted algorithms, dependent on infilling methods
Contrastive Explanations via Projection [25]	Produces explanations by projecting input representations	Focuses on causal factors, simplifies communication, alleviates cognitive burden
Model Contrastive Explanations [63]	Uses gradient-based attribution for input saliency	Measures saliency based on output, improves contrastive model simulatability
GYC [38]	Generates plausible, diverse, goal-oriented explanations	Uses Kullback–Leibler divergence for plausibility, flip labels for NLP classifiers
CAT [9]	Provides contrastive explanations using attribution classifiers	Ensures minimal perturbations, maintains semantic coherence
POLYJUICE [60]	Allows control codes for perturbations	Emphasizes grammatical correctness, minimal changes
MICE [48]	Generates minimal, fluent contrastive edits	Uses a two-stage process with fine-tuned T5 model
RELITC [8]	Generates counterfactuals by masking important tokens	Uses entropy-based strategies, produce examples close to original text

Several methods for generating counterfactual text samples are discussed (Table 1). Moreover, there has also been a significant interest in generating counterfactuals with generative models. GYC (Generate Your Counterfactual) [38] generates plausible, diverse, and goal-oriented explanations through a model-agnostic approach, employing techniques like Kullback–Leibler divergence to ensure plausibility and flipping labels for NLP classifiers. CAT (Contrastive Attributed explanations for Text) [9] provides contrastive explanations using attribution classifiers, ensuring minimal perturbations and maintaining semantic coherence while indicating topic changes. POLYJUICE [60] allows users to specify control codes for perturbations, emphasizing grammatical correctness alongside minimal changes. MICE (MInimal Contrastive Editing) [48] generates minimal, fluent contrastive edits using a two-stage process with a fine-tuned T5 model, albeit requiring significant computing power. RELITC (Relevance-based infilling for Natural Language Counterfactuals) [8], similar to MICE, generates counterfactuals by masking tokens based on importance and employs entropy-based strategies, producing examples closer to the original text efficiently. These works aim to generate minimal; one can quickly calculate grammatically sound counterfactual examples. However, such counterfactual examples are crucial in understanding how a data sample x needs to change to reach y' from y, such as in algorithmic recourse [29]. These approaches show that it is possible to generate counterfactual examples with LLMs but do not address their contrastive self-explanation abilities.

Much research has been done on LLMs in terms of counterfactual examples generation for text data [8, 9, 38, 48, 60], and self-explanations [23, 39]. Many works have also explored contrastiveness regarding gradient-based approaches in NLP [25, 63]. Even a very recent work generated CEs for LLMs [36]. However, to our knowledge, no other research has generated contrastive **self**-explanation with LLMs and evaluated them on their generality, relevance, usefulness, and readability. Unlike most works in Table 1, our work evaluates contrastive self-explanations in an automated quantitative manner and through a user study (questionnaire).

2.4 Evaluation Criteria

Self-explanation faithfulness in LLMs refers to how accurately the model's explanations reflect its actual reasoning process. Recent research has highlighted significant concerns about the reliability of these self-explanations. Madsen et al. [39] have shown that self-explanations from LLMs should not be universally trusted, as their faithfulness varies depending on the specific explanation type, model, and task. The authors noticed this is highly task- and model-dependent and requires further research to understand faithfulness better. Similarly, Agarwal et al. [2] focus on the dichotomy between plausibility and faithfulness in LLM self-explanations. An explanation can be convincing without being faithful to the model's reasoning. They also address the open challenge of enhancing faithfulness, which is currently critical in real-world applications for high-stakes decision-making. Parcalabescu et al. [44] make a significant claim regarding the nature of self-explanations provided by LLMs. Their key argument is that what was previously considered "faithfulness" in LLM explanations is better characterized as "self-consistency." For this reason, we refrain from using "faithfulness" but rather "self-consistency". Self-explanations can also be in the form of chain-of-though reasoning, token importance and counterfactual explanation. This work utilizes token importance through text redaction, extending this method to a contrastive setting (§3.3).

Yin and Neuburg [63]'s work focuses on explaining language models contrastively, where they look at input token saliency and measure the probability differences between the expected outcome (fact) and an alternative scenario (foil), thereby quantifying the degree to which specific factors influence the model's decision in a contrastive manner. Their work looks at prediction rather than text generation as they try to explain how to predict the next token. Our work does not look at probabilities of predicting the next token but rather at a contrastive explanation. We also utilize word embeddings, but not from the tokenizer that comes from the model. As NLP advanced, measuring semantic similarity between texts became quite robust. In their work, Zhang et al. [65] introduced an NLP-based approach to measuring similarity between texts using BERT-like models. This method computes the similarity of tokens using embeddings from models pre-trained on large corpora instead of simple word-to-word comparisons. These embeddings capture context, which significantly influences their meaning. This method correlates better with human judgments and demonstrates robust performance on challenging tasks like paraphrase detection. For comparison, Corley and Mihalcea [11] proposed knowledge-based algorithms to assess semantic similarity between text segments. However, contextual embeddings generally offer a more nuanced understanding of text similarity by incorporating context, which is especially valuable in cases like paraphrase detection or when evaluating translations and summaries.

Flesch-Kincaid Grade Level [16] metric measures readability based on sentence length and word syllable count. However, [54] suggests that FRGL should not be used as a metric, as it can be easily manipulated. They suggest that only the FKGL components are reported instead, i.e., average sentence length and the average number of syllables, which demystifies the readability score and provides concrete information about the types of changes that are being made by the systems. Recently, Rooein et al. [47] proposed prompt-based metrics derived from user studies that involve asking LLMs specific questions about text features. They capture more abstract aspects of text difficulty, such as educational level and lexical complexity, and have been shown to improve text difficulty classification when combined with traditional metrics.

Evaluating contrastive self-explanations intelligibility, relevance, and usefulness almost invariably requires human feedback. For domains requiring specific knowledge (e.g., medicine and law), engaging domain specialists to assess the quality of the generated text is crucial. This approach is more reliable than crowdsourced evaluations, especially for specialized or creative content [30]. Ikart et al. [24] inform us that expert reviews can identify significant problems with questionnaire design early in the development process before extensive resources are invested. Therefore, it is worth ensuring the questionnaire content is well-constructed. However, engaging experts is generally more costly and less scalable than crowdsourcing. This can be a significant barrier, especially for large-scale projects [10], where many responses are needed, and the budget is not large enough to cover the costs of employing and finding such experts.

3 Methodology

The methodology starts with task selection, which defines the requirements for a dataset structure. Secondly, we go over the LLMs we chose for our method. We evaluated their accuracy on the selected task to determine whether they were accurate in their prediction and were good candidates for generating contrastive explanations. Moreover, we verified whether the models could stay self-consistent in responding to our prompts. That led us to explanation generation, where we generated contrastive and non-contrastive explanations. We then evaluated the explanations using an automated approach and a questionnaire involving participants. The section introduces all the steps required to generate and evaluate contrastive self-explanations.

We proposed a new method called the **Contrastive Self-Explanation Method (CoSEM)**, which generates and evaluates instruction-tuned LLM contrastive self-explanations (Figure 3). As far as we know, CoSEM is the first method that generates contrastive self-explanations for LLMs and evaluates them automatically and through a questionnaire. The method consists of five steps. First, *task selection* (§3.1) adjusts the dataset for generating self-explanations. *Model accuracy evaluation* (§3.2) ensures that the model used for generating those self-explanations is capable of correctly distinguishing between the fact and the foil. *Model self-consistency verification for CEs* (§3.3) is an extended method from [39], which helps us verify self-consistency in LLMs for CEs. It aims to ensure the model is not blatantly hallucinating, although it is a vast known drawback in LLMs [62]. The crucial part of CoSEM is *explanation generation* (§3.4). It generates four types of explanations: contrastive self-explanations, modified contrastive self-explanations, and non-contrastive self-explanations. This part of the method also uses a novel rule-based syntactic modification approach to generate generic and semantically similar text. This sub-method helps us generate generic modified contrastive explanations for evaluating the generality of contrastive self-explanations. Finally, we evaluate the generated contrastive self-explanations using two approaches: automated *explanations evaluation*(§3.5) and *questionnaire* (§3.6). The first approach uses automation to evaluate the explanations for usefulness, generality, and readability. The second approach deploys a questionnaire to evaluate self-explanations of usefulness, intelligibility, and relevance. We included all the detailed step descriptions in the following subsections. For completeness, two examples of prompts for the movie dataset and the medical dataset can be found in Appendix D.1 and Appendix D.2, respectively.

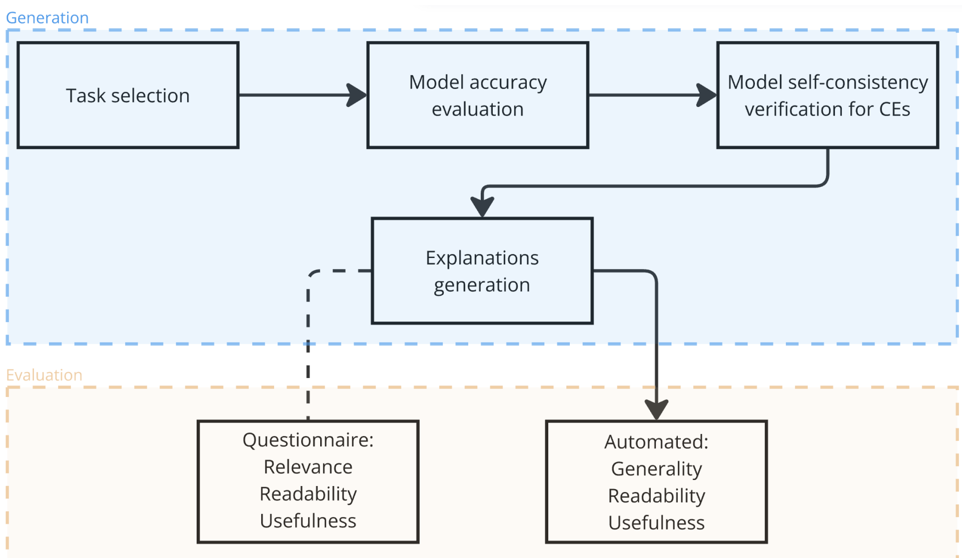


Figure 3: Flowchart representing the overview of Contrastive Self-Explanation Method (CoSEM)

3.1 Task Selection

In the initial step of our methodology, we focused on task selection to ensure that the chosen task is suitable for evaluating the ability of LLMs to generate contrastive self-explanations. The selected task is *text classification*. This choice is apparent when contrastive explanations refer to fact and foil, essentially two opposite classes to which we can classify a text. However, since some datasets (as seen later) consisted of multiple labels, we narrowed text classification datasets to binary classification. In short, the correct label for a given text in a dataset is the fact, while an incorrect class is a foil. When a contrastive question is asked, the questioner assumes that the fact is accurate and that the foil is false [58]. Although the contrast case (foil) can have more than one alternative, we narrowed it down to one [57]. In case of multiple incorrect classes, we select one that is least semantically similar (§3.4.1) to the fact. We do this selection to increase the contrast between the fact and the foil as much as possible.

Although we utilize text classification for the task, the method can be expanded to other tasks, such as question-answering. The authors of [39] perform their faithfulness verification on sentiment classification, multi-choice classification, and two-paragraph classification. Those tasks can be adapted to our approach in the task selection step. For example, question-answering is a multiple-choice question with four answers (one fact and three foils), as well as natural language inference (NLI) (one fact and two foils). However, these extensions could be investigated in future research.

Task selection also involves identifying relevant datasets and determining the specific classification tasks that will allow us to assess the model's explanation capabilities effectively. We have selected two diverse datasets: *the Medical-Abstracts-TC-Corpus dataset* [50] and *the IMDB reviews dataset* [37].

Medical Abstracts TC Corpus The Medical-Abstracts-TC-Corpus [50] contains medical abstracts detailing the current conditions of patients, categorized into five distinct classes: *digestive system diseases*, *cardiovascular diseases*, *neoplasms*, *nervous system diseases*, and *general pathological conditions*. We chose this dataset for a few reasons. First, it contains domain-specific vocabulary and topics which might be challenging for an average respondent to understand. For this reason, we wanted to evaluate the models' capabilities in giving contrastive explanations to see how well they can explain topics. There is a chance that smaller models might not be trained on such specific data, which is not common knowledge. The dataset includes a range of medical conditions, providing a broad spectrum of topics for the LLM to generate explanations. This diversity helps evaluate the model's ability to handle various domains within the medical field. Secondly, the medical domain's real-world relevance and the critical nature of accurate information make this dataset ideal for testing the robustness and generality of the LLM's explanations. Ensuring the model can provide reliable explanations in a high-stakes domain like healthcare.

Internet Movie Database (IMDB) The IMDB reviews dataset is used for binary sentiment classification and consists of highly polar movie reviews. Additionally, a substantial amount of unlabeled data is available for further analysis. This dataset is of interest due to a few reasons. First, the binary nature of sentiment classification (positive or negative) provides a straightforward task for generating contrastive explanations. The model must explain why a review is classified as positive rather than negative or vice versa, making it an excellent test case for contrastive self-explanation capabilities. Moreover, movie reviews can vary significantly in style, tone, and content, providing a diverse set of texts for the model. This variability is essential for testing the model's ability to generate consistent and contextually appropriate explanations across different text types.

3.2 Model Accuracy Evaluation

First, we evaluated whether the model is suitable for CE generation. If the model cannot correctly predict the fact label for a text sample, it will falsely try to generate a self-explanation. Given a text sample, we prompted the model to predict what label it belongs to. Figure 4 shows an example of the evaluation.

Algorithm 1 Predicting the label of a text with an LLM

```
function PredictLabelLLM(text, fact, foil)
    Input: Text sample text, the fact label fact, the foil label foil
    Output: Model prediction, either fact or foil. If the model cannot predict, then "unsure".
    response ← LLM(prompt)
    if response contains only fact then
        return fact
    else if response contains only foil then
        return foil
    else
        return "unsure"
    end if
end function
```

What label does the following text belong to: $\{fact\}$ or $\{foil\}$? If you can't tell, then "unsure". Output only one word.

$\{text\}$

$\{response\}$

Figure 4: Example of prompting the model to predict a sentiment (label) for a text sample. The model responds with either positive, negative or unsure. If the response is neither of those, then it is "unsure".

In the example above (Figure 4), the model is explicitly asked to output only the word corresponding to the label (e.g., positive, negative or unsure). The "unsure" option is used in the prompt later, in the verification phase (§3.3), and for consistency purposes across prompts, it is kept the same.

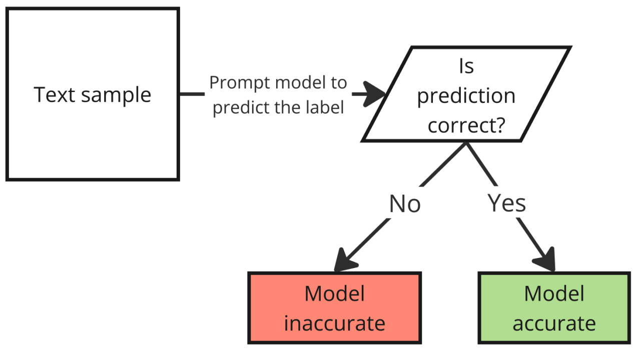


Figure 5: Flowchart representing the evaluation of model accuracy

If the model cannot stay accurate, we do not proceed further with the sample. We move on to the following sample and mark this as "inaccurate".

3.3 Model Self-consistency Verification for Contrastive Explanations

Before we generated contrastive explanations, we verified that we can trust the model in what self-explanations it generates for us. The initial step in establishing whether a model can do that is to check its self-consistency.

Parcalabescu et al. [44] note that we cannot claim that a model is faithful if it is self-consistent, as it is insufficient for faithfulness. We leave this part for future research as this paper focuses on the generality of contrastive self-explanations.

The original work compares three self-explanation methods: counterfactual, token importance and redaction. In essence, the latter two are similar, where we replace tokens with a mask token. The counterfactual method, however, prompts the model to edit a text so that the fact label turns into foil. This approach reduced the amount of work required compared to token importance (i.e., prompting the model to find required tokens, then replace them manually) and also had better success on the text classification task in the original method paper.

We extended a part of the work done by Madsen et al. [39], where they perform model self-consistency using redaction. Redaction is a method in which a model is prompted to modify an input text by removing tokens (redacting with a mask token) that would prevent it from predicting what the text is about. We adjusted the method in which the model redacts the words needed to generate a contrastive explanation. The method aims to evaluate whether a model is self-consistent in its self-explanations, so we extend it to a contrastive setting. Initially, we wanted to utilize the counterfactual method, as it performed better on task classification in their work. However, to extend it to a contrastive setting, we prompt the model to redact tokens necessary to generate a contrastive explanation. In Figure 6, we start the method by generating a contrastive explanation to ensure that it is possible with the model and to prompt the model for a redaction (Figure 38) in the same session. Moreover, the original paper justifies the choice of masking tokens, where the authors state that classification should be robust to [REDACTED] tokens. Similarly, the authors also explored the use of voice in prompts, where they address the model (1) objectively ("What is the..."), (2) by referring to "you" ("What would you..."), or (3) addressing a human ("What would a human..."). There was no significant preference towards any of the voices. For this reason, we stayed consistent with the use of voice: objective.

We performed multiple model self-consistency checks to ensure the approach was quantifiable. Figure 6 displays the process of self-consistency. The models we use have a controllable parameter `seed`, which controls the sampling process, ensuring varied outputs for each model. Using the same seed guarantees the same output. We performed redaction on $k = 5$ different seeds to generate five different redacted texts. Choosing high k was problematic time-wise—a large amount of time was needed for each model initialization (with a new seed) and for inference. We generated k different responses using each model for every text sample. We used each sample as a prompt for the model to classify the sample. If the model couldn't classify the sample (it responded with "unsure"), it was self-consistent (it masked words that would prevent it from knowing how to classify the text). If it classified the sample, then it was not self-consistent. Based on the k samples for each text, we took an aggregate score of successful and failed classifications.

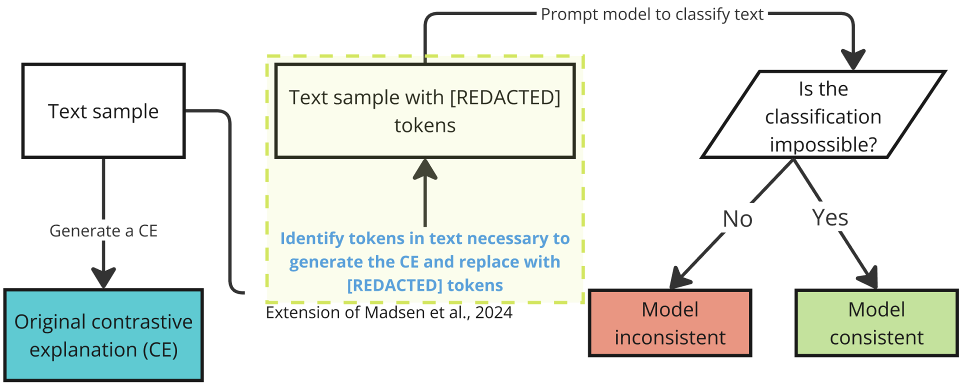


Figure 6: Flowchart for model self-consistency verification for CEs. The method extends an existing work done by [39]. The light green box in the middle indicates the part creating multiple redacted texts with various model seeds.

> For the following text, why is it $\{fact\}$ and not $\{foil\}$?
>
> $\{text\}$

> $\{contrastive\ explanation\ response\}$

Figure 7: Example of prompting the model to generate a contrastive explanation ("Why the fact and not the foil?")

Next, instead of finding words/phrases needed for sentiment prediction, we found the ones required to generate the CE. The prompt for that is visible in Figure 8. We asked the model to explicitly give us all the words/phrases needed.

> In the text below, redact the words necessary to answer question why {fact} and not {foil}. Replace important words with [REDACTED]. Do not explain the answer. Give only redacted text.
>
> $\{text\}$

> $\{redacted\ text\ response\}$

Figure 8: Example of prompting the model to identify necessary tokens for reaching the contrastive explanation. The method has been adapted from Madsen et al. [39].

Finally, we prompt the model to predict the label of the sentiment of the text sample with replaced tokens, using the same prompt as in Figure 4 but by replacing the text with the redacted text. If the model can make a prediction based on the text, it is self-inconsistent, as according to the model itself, we removed all the tokens needed to create one (Figure 38). However, if the model did output "unsure", we accepted it as a self-consistent model for contrastive explanations.

> What is the class of the following text? The text can contain redacted words marked with [REDACTED]. Answer only {fact}, {foil}, or unsure. Do not explain the answer.
>
> {*text*}

> {*predicted class*}

Figure 9: Example of prompting the model to predict the text on the redacted text. If the model predicts "unsure", then it is faithful. In any other case, it is not faithful.

3.4 Explanations generation

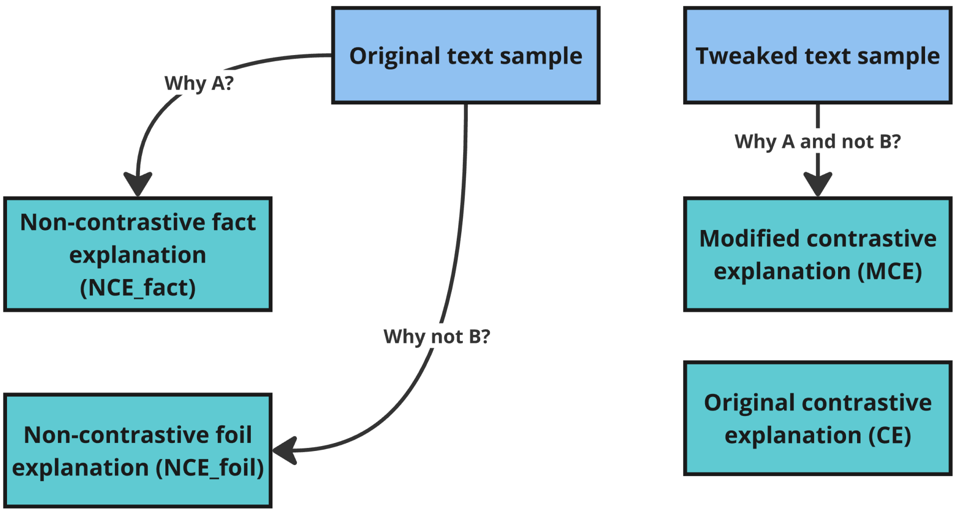


Figure 10: Flowchart for generating explanations step. We retrieve the CE from the self-consistency step performed earlier. Non-contrastive explanations ($\mathrm{NCE}_{\mathrm{fact}}$ and $\mathrm{NCE}_{\mathrm{foil}}$) are generated from the original text, while modified contrastive explanation (MCE) comes from a modified sample.

The final step before the explanations evaluation is their generation. We have generated a contrastive explanation in §3.3. We further generated two non-contrastive explanations for the fact label ($\mathrm{NCE}_{\mathrm{fact}}$) and the foil label ($\mathrm{NCE}_{\mathrm{foil}}$). We prompt the model to generate the explanations by asking non-contrastive questions: "Why *fact*?" and "Why not *foil*?" respectively (Figure 11). We also generated a modified contrastive explanation (MCE). This is a contrastive explanation generated based on a modified text, which we described in the following subsection (§3.4.1)

For the following text, why is it $\{fact\}$?

$\{text\}$

$\{non-contrastive\ explanation\ (fact)\ response\}$

For the following text, why is it NOT $\{foil\}$?

$\{text\}$

$\{non-contrastive\ explanation\ (foil)\ response\}$

Figure 11: Example of prompting the model to generate non-contrastive self-explanations (NCEs) for the fact and the foil labels.

3.4.1 Generating Modified Contrastive Explanation (MCE) using Rule-Based Syntactic Modifications

To evaluate generality, we needed to modify the input text syntactically so that its meaning would not be affected. We found a way to provide a model with a modified version of the input text based on which the model generates an explanation. The text modification then allows us to ensure that the text stays semantically close to the original. If models generate a general contrastive self-explanation (semantically similar to the original contrastive explanation) regardless of how syntactically modified the text is, they can generate explanations for this topic. We looked at hypernyms (words with a broader meaning than the target word) and synonyms (words similar to the target word) as those word replacements do not change the semantic sentence meaning.

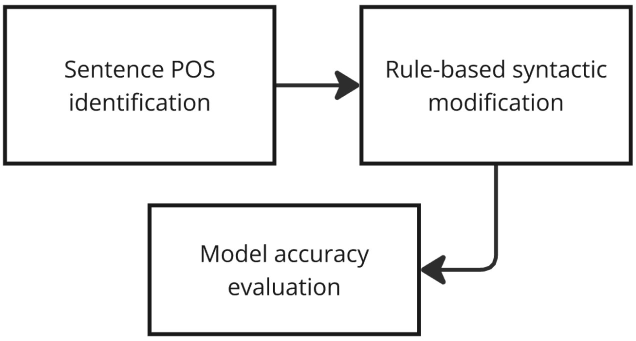


Figure 12: Flowchart for contrastive explanation tweaking. This approach consists of three stages: identifying target POS, replacing them with the closest hypernyms or synonyms, and verifying that the label does not flip.

The following method consists of three steps (Figure 12):

Sentence POS Identification The first step in performing syntactic modifications is to identify the sentence's components, such as nouns and adjectives. Stanza NLP library [18] is used to modify the text, which provides detailed part-of-speech (POS) tagging for each word in the sentence. The library helps determine whether replacement words are of the same POS (Algorithm 3).

Rule-based Syntactic Modification The identified components (nouns and adjectives) are subject to rule-based syntactic changes. Specifically, we replaced these components with their hypernyms and synonyms. The goal is to slightly tweak the text while preserving its original meaning (Algorithm 3). The algorithm retrieves the top-k hypernyms and synonyms for each identified component using the WordNet lexical database. We obtained each word's embeddings and hypernyms/synonyms using the `bert-based-uncased` model and computed cosine similarity to find the closest replacement that maintains semantic similarity. The closest hypernym or synonym is selected based on the highest cosine similarity score. The selected hypernyms or synonyms replace the original words in the text, resulting in a slightly modified version of the original sentence.

Model Accuracy Evaluation Once we generate a modified text sample, we prompt the model to classify its sentiment/label to confirm the new text sample is also of the same sentiment/label.

Algorithm 2 Rule-based Syntactic Modification. This procedure takes the original text as input and uses Stanza to tokenize the input (*ParseWithStanza*) to identify all nouns and adjectives, then tweak them (*TweakWord*).

```
function ModifySyntactically(text)
    Input: Original text sample
    Output: Tweaked text sample with modified nouns and adjectives.
    sentences ← ParseWithStanza(text)
    for all sentence in sentences do
        for all word in sentence do
            if POS(word) ∈ ["NOUN", "ADJ"] then
                word ← TweakWord(word)
            end if
        end for
    end for
    return sentences
end function
```

Using WordNet's extensive database of lexical relationships [15], particularly synonyms and hypernyms, is an effective strategy for modifying text while maintaining semantic similarity and achieving a more generic tone. WordNet's synonyms allow replacing words with others with the same or nearly identical meaning, ensuring that the modified text retains its original intent and context. Conversely, hypernyms provide a hierarchical structure that enables substituting specific terms with broader categories. This approach helps maintain the text's coherence and generality and ensures that the modifications do not introduce unintended nuances or specificities. Nadig et al. [43] highlighted the robustness of WordNet's structure and its validated relationships, making it a reliable resource for NLP tasks. By utilizing these relationships, text modification can balance semantic fidelity and generalization, which is crucial for tasks that require maintaining the essence of the content while broadening its applicability. This approach aligns well with the necessity for producing accurate and broadly understandable explanations, thereby enhancing the utility and accessibility of the generated text.

Algorithm 3 Word replacement. Here, the closest k hypernyms (and synonyms) are found based on Cosine similarity between word embeddings using `bert-base-uncased` model.

```
function TweakWord(word)
    Input: Original word word
    Output: Modified word
    closest_hypernym ← FindClosestHypernym(word, k)
    if closest_hypernym exists then
        return closest_hypernym
    else
        closest_synonym ← FindClosestSynonym(word, k)
        if closest_synonym exists then
            return closest_synonym
        else
            return word
        end if
    end if
end function
```

The `TweakWord` (Algorithm 3) is designed to modify a given word by finding its closest hypernym or synonym based on cosine similarity between word embeddings. This method utilizes the `bert-base-uncased` model to compute the embeddings and identify the most semantically similar words. The algorithm aims to replace the original word with a more generic (hypernym) or equivalent (synonym) term while maintaining the text's contextual integrity. This approach ensures that the replacement word maintains a close semantic relationship with the original word, enhancing the text's generality and applicability while preserving its meaning.

Closest hypernym/synonym We retrieved a word's top k hypernyms/synonyms, then found semantic similarity and compared with cosine similarity with the original word to get the closest. We disregard scores of 1.0 or higher as, in such cases, the hypernym is the same word as the original. We use the $k = 0.7$ threshold to cut off any hypernyms/synonyms that are not semantically similar.

3.5 Evaluating Model Self-explanations

In this section, we performed three automated evaluations: one comprises evaluating usefulness (§3.5.1), the second generality (§3.5.2), and the last one explanation readability (§3.5.3). We, therefore, utilized the four explanations generated in our three distinct evaluations.

Semantic Text Similarity An important aspect we utilized in this work is the similarity between texts. Commonly, we might wonder how two texts relate by checking how much overlap there is between them. However, we wanted to go beyond lexical matching. We utilized Semantic Text Similarity (STS), which aims to capture the underlying meaning of texts rather than just matching words. This approach allowed us to identify similarities between explanations even when different words or phrasings express the same concept.

Semantic similarity is addressed through text tokenization and calculating word embeddings (for each word). The word embeddings are vectors of numbers, which can be used to calculate a numerical distance. The closer the embeddings, the more semantically similar they are. This approach works great for comparing two words to each other and finding how semantically close they are. Nevertheless, suitable STS methods are required since lengthy explanations comprise hundreds of words. BERTScore [65] and Sentence-BERT [46] are two, as far as we are aware, state-of-the-art approaches for comparing the semantic similarity between texts (not single words) and are commonly used in STS tasks.

We utilized STS in evaluating usefulness using BERTScore (3.5.1) and generality using Sentence-BERT (3.5.2).

3.5.1 Usefulness

Semantic Similarity BERTScore uses contextual embeddings from BERT for each token in the compared sentences and employs a greedy matching strategy to find the best matches between tokens in the two sentences. It calculates precision and recall, encapsulated with a third metric - the F1 score. It also does not require fine-tuning for the similarity task, but it tends to produce higher similarity scores, even for dissimilar sentences, which can be misleading in some cases. BERTScore compares contrastive explanation (paragraph) with a label (1-3 words). It is adequate for tasks where fine-grained token-level similarity is essential, which is not valid with Sentence-BERT. We used BERTScore to ensure the label is thoroughly checked against all the text in the contrastive explanations for semantic similarity. Using semantic similarity to compare prompts, Luss et al. [36] use semantic similarity. However, they do not disclose the reason for using the `google-bert/bert-base-uncased` model for computing the word embeddings. This model is widely used in the NLP community, making it a familiar choice for many researchers and practitioners. It is significantly smaller and less resource-intensive, hence more appropriate for processing extensive data. BERT is also trained on a large corpus of uncased English text and provides robust general-purpose embeddings that are effective across a wide range of tasks and domains [13]. However, larger and more enhanced models such as `microsoft/deberta-xlarge-mnli` [20] have better correlation with human evaluation [65]. For this reason, in our evaluations, we utilize `microsoft/deberta-xlarge-mnli` for Sentence-BERT.

Evaluation This evaluation compared the semantic overlap between the label and the explanations. The purpose was to establish whether contrastive explanations are semantically closer to either label, implying the models' preference when generating such explanations. Furthermore, we also compared the non-contrastive explanations to their labels (i.e., NCE_{fact} to fact and NCE_{foil} to foil) and the opposing/unrelated label. NCE should be explaining its label; therefore, it is more likely to be semantically similar to it rather than the unrelated label. We used this comparison as the baseline for comparing the contrastive explanations. For example, NCE_{fact} should be more semantically similar to fact than to foil. Conversely, contrastive explanations should be semantically similar to both labels, more or less equally, as a contrastive explanation should address both the fact and the contrast. We have utilized the BERTScore method to compute F1 scores between the explanations and the labels.

BERTScore calculates the F1 score by combining precision and recall metrics based on semantic similarity between reference (text) and candidate (label) texts. The precision is calculated as follows: for each token in the candidate sentence, it finds the most similar token in the reference sentence, takes the maximum similarity scores, and then averages these scores. Conversely, the recall is calculated similarly, but candidate and reference sentences are swapped. Finally, the F1 score is the harmonic mean of precision and recall, i.e., $\frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$. For example, if the F1 score of the CE-fact is higher than the CE-foil, then the CE is semantically more similar to the fact label.

3.5.2 Generality

Similarity Similarity Sentence-BERT is an extension of BERTScore. It creates a single embedding vector for the entire sentence rather than a single word. It achieves this by pooling the token embeddings. It then calculates the distance between the sentence embeddings rather than tokens. The approach is optimized for this task [21], where we must understand the overall meaning of paragraphs. This is also the reason we chose Sentence-BERT. We used an existing implementation for Python, Sentence-BERT [1]. It gives an option for selecting a sentence-transformer model. This model is responsible for tokenizing the words used to compare the sentences. Among many options, we have opted for a pre-trained `sentence-transformers/all-mpnet-base-v2`[2] model, which has high performance on sentence embeddings and semantic search. We have considered an alternative (default option provided by the library) - `sentence-transformers/all-MiniLM-L6-v2`[3]. However, that model only supports paragraphs no longer than 256 words. On the other hand, the model we selected supports up to 384 words. The only drawback is its longer computation time.

Evaluation This part evaluated the model's ability to generalize explanations despite lexically different inputs. We believe that more general explanations give a better picture to capture the information that is being explained. We assessed the generality of the explanations by comparing a contrastive self-explanation to its modified version. Audemard et al. [4] note that providing many contrastive explanations to respondents is often impractical, as they may be overwhelmed to comprehend them all. Conversely, reducing contrastive explanations to single instances can result in highlighting outliers rather than genuine contrastive explanations. Therefore, offering a single explanation that captures the general information is more effective.

Lexical diversity and density Generality implies that we look at common traits that are not specific to an instance. We measured lexical diversity and density to check for text generality and, simultaneously, the complexity of words used within it. First, we measured the *lexical diversity*, which measures the number of different words used in both contrastive explanations and provides a measure of the proportion of POS in them [7]. The metric provides insights into how varied the language in the explanations is. The lower the difference between the two, the more general the explanations are. A standard metric for lexical diversity is the type-token ratio (TTR), which measures the ratio of different words (types) to the total number of words (tokens). However, its drawback is that longer texts generally have lower TTR (as the number of words increases "infinitely") [28]. Using it is only reasonable when two texts are of equal length, which is rarely true for LLM-generated explanations. To measure the lexical diversity, we utilized the Moving-Average Type-Token Ratio (MATTR) proposed by Covington et al. [12]. It is a text-length independent adaptation of TTR. It is a preferred method for comparing text ranging from 50 to 200 words [64], which is the range of our explanations. There is no one valid way to calculate it, as the window depends on a combination of empirical testing and theoretical consideration of the linguistic features of the explanations we are comparing. We selected a window of 50 words, the lower bound of the preferred text size. We used a Python library, LexicalRichness [49], to calculate word counts and MATTR. On the other hand, we calculated *lexical density*, which is "the proportion of content words (nouns, verbs, adjectives, and often also adverbs) to the total number of words" [28, p. 65]. Higher density suggests that the text is more functional or descriptive. This evaluation determines whether one explanation is more general or specific.

We used Sentence-BERT to compare two paragraphs for their semantic similarity. Sentence-BERT should return a high cosine similarity score between two explanations if the model can generate them at a high generality, implying that the explanations are not very different from e ach other.

[1] https://www.sbert.net/
[2] https://www.sbert.net/docs/sentence_transformer/pretrained_models.html
[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

3.5.3 Readability

Apart from being contrastive and general, we believe the explanations should be easy to read. When that is the case, the contrasts become more apparent and memorable due to lower cognitive load [53]. Given that LLMs are accessible to the public, it is only valid to assume that the respondents' reading levels can be at various levels. For this reason, we evaluated the explanations according to the Flesch-Kincaid Grade Level (FKGL) [16]. It is a straightforward metric that combines two text statistics to calculate the score: the average number of syllables per word and the average number of words per sentence. However, [54] suggests that FKGL should not be used as a metric, as one can easily manipulate it. Nevertheless, they state that using its two components is preferred. We, therefore, calculated the scores for the two components of FKGL (average number of syllables per word and average number of words per sentence). We compared the metrics of explanations to the original text, which we used as the baseline. The lower the numbers, the easier the text is to read [16].

Average Number of Syllables Per Word We first tokenized the text into sentences and tokens using the Stanza (similarly as in §3.4.1). For each token, we counted the number of syllables using a Python library `hyphenate`[4], which is based on Frank Liang's algorithm [32] for word hyphenation. We then averaged that number for the whole text. Ultimately, we calculated the average number for all texts and reported the standard deviation.

Average Number of Words Per Sentence This number was straightforward - we again utilized the Stanza to tokenize the text and retrieve sentences. We then counted the number of words per sentence we averaged per text. Finally, we averaged that number across all texts and reported the standard deviation.

Average Words Per Text Additionally, we decided to compare the length of the explanations. That is a more generic metric that gives us a view of how the explanation is compared to the original text. We noticed that some texts are longer than others, and hence, it is not necessarily fair to not acknowledge that.

Limitation A limitation that we have faced is the POS tagging and tokenization. The tokenization considers punctuation symbols (non-English words) as actual words (which have syllables). As an example, a dot (.) is considered as a 1-syllable word. We decided not to filter those out for two reasons. First, it required additional rule-based preprocessing and would have caused additional processes to establish what rules to use (e.g., keep only the content words that are relevant to the context). Second, our comparison of explanations to text is relative. Therefore, if we keep the method consistent for both the explanations and the text, we are treating them the same and look only at the difference between the two.

Assumptions We have made a few assumptions regarding the readability of the explanations and the text. First, we assumed that the task topic reflects the text complexity, meaning the medical dataset should have higher metric scores for syllables per word and words per sentence (except for maybe the words per text, as that is not necessarily dependent on complexity). We also hypothesized that contrastive explanations are more straightforward to understand due to lower cognitive load (as stated in the literature).

3.6 Questionnaire: Evaluating Contrastive Self-Explanation Intelligibility, Usefulness, and Relevance

We deployed a user study through a questionnaire to evaluate the effectiveness of LLM-generated contrastive self-explanations. The questionnaire evaluated the usefulness, relevance, and intelligibility of contrastive self-explanations generated by these language models. Measuring these three features of CEs inherently involves subjective assessment, as these qualities are intrinsically linked to individual human perception and understanding. Unlike objective measures, which can be quantified and verified independently of personal feelings (such as accuracy rates in classification tasks), assessing a CE's effectiveness relies on the individual and contextual experiences of the user [31]. Previous studies noticed that a shorter survey utilizing a brief questionnaire was reliable and produced higher response and completion rates than a long survey [51]. Moreover, when respondents lose interest due to the length of a questionnaire or the extensiveness of interviews, they tend to provide thoughtless and unreliable answers. That is precisely why our questionnaire is short and contains a brief question for each CE feature. We got inspired by Hoffman et al. [22], where in their study, they evaluated AI systems being explained to them, utilizing human feedback also through a questionnaire. Similarly, we selected

[4] https://pypi.org/project/hyphenate/

a five-point Likert scale (compared to a seven-point scale) because it increases response rate and quality and reduces respondents' frustration level [3]. However, they do not explain their reasoning for using a scale starting with 5, "I agree strongly," and ending with 1, "I disagree strongly." We believe that a more intuitive approach is to start with 1 and finish with 5.

Usefulness Starting with *usefulness*, we aimed to check whether a generated CE is beneficial to the person who seeks an explanation in distinguishing between the two contrasts (Appendix A). As contrastive explanations should address both the fact and the foil, we expected respondents to receive an explanation explaining the contrast between them. To evaluate that, we provided the respondents with CE and asked them how helpful the contrastive explanation is in distinguishing between the two labels. To measure the responses, we utilized the Likert scale [33], a technique used to acquire quantified data from questionnaire responses. When creating a scale, general recommendations are to use clean constructs, make readable items, and choose an appropriate number of points to ensure clear and distinguishable response options. Our questionnaire was crafted with due consideration of these recommendations by [26].

Readability Furthermore, we evaluated *readability* through a questionnaire (Appendix A), additionally to the automatic lexical evaluation. Evaluating readability through a user study is crucial to ensure that a contrastive self-explanation generated by an LLM has a purpose and can be used by the person reading it. If an explanation is not readable, then it is not meaningful to the user and hence not practical. We performed this user evaluation using a five-point Likert scale.

Relevance Finally, we evaluated the *relevance* of CEs w.r.t. to the original text (Appendix A). The relevance is the quality of being closely connected. We evaluated whether the contrastive explanation is closely connected (in other words, appropriate) to the original text and, hence, relevant to the respondent. LLMs use contextual information to generate their responses, even if they are not relevant to the task [59]. Therefore, a contrastive self-explanation that appears relevant according to a respondent might not be factually correct. However, we focused on evaluating CEs based on the context we give to an LLM (by providing the text in our questionnaire for reference).

Selecting Questionnaire Samples Due to limited resources and time, we could not perform an extensive crowd-sourced questionnaire to evaluate multiple samples from each dataset. The limitations of this study are further described in Section 6. Due to limited resources and time, we conducted a shorter questionnaire. We collected 20 and 10 responses for the Movie reviews and the Medical abstracts datasets, respectively. We selected one sample from each dataset, as we could not afford to evaluate more. It was challenging, as each sample could yield different results from the respondents. Therefore, we introduced criteria based on which we narrowed down the choice of text and the five explanations. There are two criteria points we considered in this particular order:

1. The sample was correctly classified by the model.
2. We picked the average length of text from the available candidates.

The first criterion aligned with our previous choice to consider only samples correctly classified by the models. The second criterion ensured the representation of an average sample from the dataset. This is not an ideal solution, but we deemed conducting a user study based on a single sample feasible. All questionnaire questions are in Section A. Appendix A.2 contains a concrete example of a questionnaire for the medical dataset, and Appendix A.3 for the movie reviews dataset.

3.7 Experiment Setup

Choosing Models We focused on instruction-tuned LLMs, ready to be used in chat/conversation settings. There are many popular models, such as ChatGPT, available online. However, some of them are not open-source. Open-source models allow us to use them locally and have control over their parameters, which is necessary for our research. These models are used for text generation, which is how they generate self-explanations. The models are publicly accessible at HuggingFace[5]. We ran the models on DelftBlue[6], which is TU Delft's supercomputer. We had access to the education account, which has relatively restrictive limits. We used

[5] https://huggingface.co/
[6] https://www.tudelft.nl/dhpc/system

DelftBlue's GPU, which is the NVIDIA Tesla A100 80GB. Otherwise, running the models (especially the more extensive versions with 70B parameters) locally would not have been possible.

Only open-access models are utilized purely for liability reasons and study reproduction purposes [39]. Moreover, all models have a fixed seed (default seed set to 0) when generating responses to make the results reproducible, similarly as in [39]. The following models are selected:

- **Llama2** 7B (`Llama-2-8B-chat-hf`)
- **Llama3** 8B and 70B (`Meta-Llama-3-8B-Instruct`, `Meta-Llama-70B-Instruct`)
- **Mistral** 7.24B and 46.7B (`mistralai/Mistral-7B-Instruct-v0.1`, `mistralai/Mixtral-8x7B-Instruct-v0.1`)

We have selected well-performing chat models as in [39]. However, as Llama3 models came out recently, we have also decided to use them to compare them with their predecessors, Llama2 models. Moreover, we did not use Falcon models in the end, as we could not set up the prompting for our needs, and due to time constraints, we decided to leave them out. All the models can be accessed and downloaded from the HuggingFace. We pick two versions of each model, one with a smaller parameter space and the other with a larger one, to compare how much more effective larger models are.

Model Parameters We set the model parameters similarly to those in [39]. More specifically, we set the default seed of 0 to all models. The only exception is when we generated redacted texts (§3.3) to get various responses (in a controlled environment). For example, if we generate five additional samples, the new use seeds numbers 1-5. We do that to generate redacted texts systematically, let the model predict with the default seed, and quantify the obtained results. We also use new model instances to prevent the model from using context in the previous prompts, which would result in a failed self-consistency check. Starting a fresh model session would effectively reset the context window.

In their work, Madsen et al. [39] set model parameters according to how the online chat tools have it. The settings are the same as in HuggingFace's online chat service[7]. However, we have opted for a different approach that guarantees reproducibility. We do not set any parameters that do not guarantee full reproducibility. They do not discuss these parameters in depth in their work, but we believe understanding how that works is essential. Firstly, `temperature` is a parameter that controls the randomness of predictions by scaling the logits before applying softmax. A higher temperature increases diversity, while a lower temperature makes the model more confident but less diverse. As the temperature parameter is higher, the LLMs moderately generate more novel outputs [45]. Secondly, `repetition_penalty` parameter reduces the probability of generating tokens that recently appeared in the generated text. Lastly, `top_k` limits the sampling pool to the top k, most likely the following words. In contrast, `top_p` instead of cutting off at the top k, it chooses from the most miniature possible set of words whose cumulative probability exceeds the threshold p. Higher `k` increases randomness, focusing on a probable subset of predictions, balancing diversity and reliability.

[7] https://huggingface.co/chat/

4 Results

Accuracy Figure 13 showed the performance achieved by the five models on the IMDB (blue bars on the left) and the Medical datasets (green bars on the right). On average, the models classified text samples with higher accuracy on the IMDB dataset. `Mistral-7B` had the words performance on both datasets, with 74% on the movie reviews and 22% on the medical diagnoses. The best-performing model was `Llama2-7B`, while a more complex model `Llama3-70B` performed just as well or worse. `Llama3-8B`, similar complexity to `Llama2-7B`, did not perform as well as its predecessor.

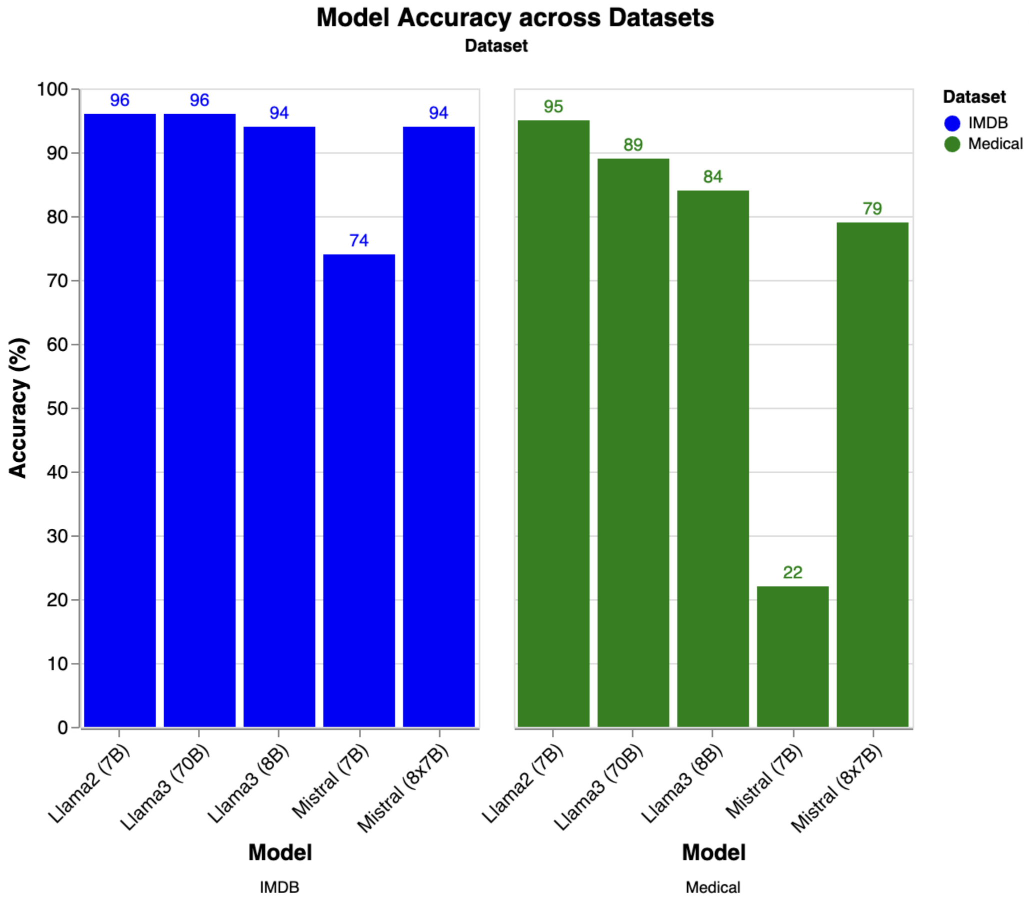
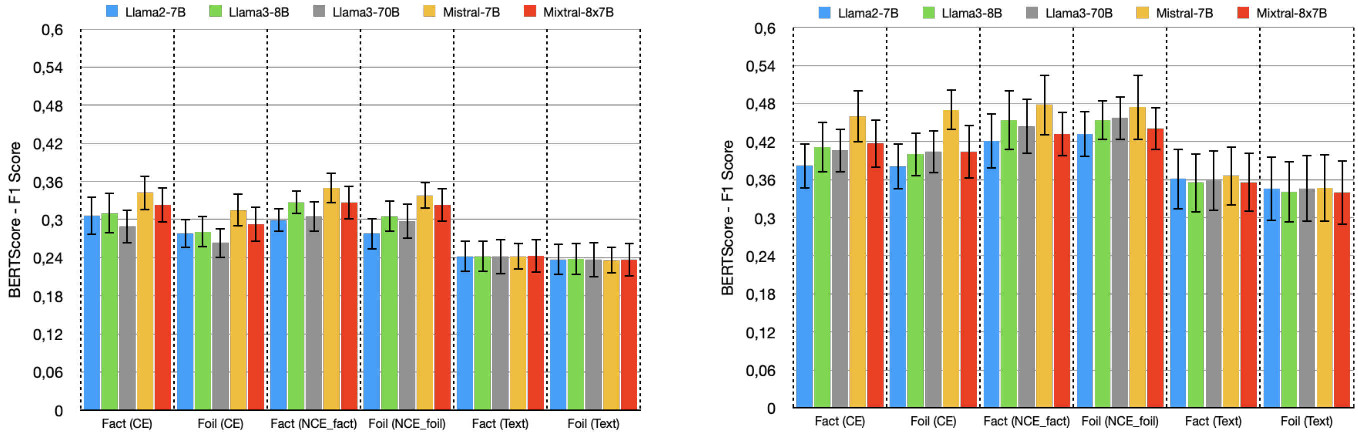


Figure 13: Models accuracy in predicting the label (fact/foil/"unsure") for all five models. All models were evaluated on 100 samples. The blue bars (left) show scores achieved on the IMDB dataset, while the bard (right) show scores achieved on the Medical dataset.

4.1 Usefulness

Figure 14a and Figure 14b show the F1 score calculated using BERTScore method, used for texts and explanations to the fact and foil labels. Figure 14a contains scores obtained from the IMDB dataset, and Figure 14b scores from the Medical dataset. The plots are subdivided into groups (separated by vertical dashed lines), representing what the two components are to each other. The bars are represented by different colors, which correspond to the five models we used in our evaluation.

The results for usefulness based on the F1 scores showed us the following. On average, the F1 scores for the Medical dataset (Figure 14b) are higher than the scores for the IMDB dataset (Figure 14b). Moreover, the last two plot groups compare text-to-fact and text-to-foil, respectively. The similarity of text-to-foil is higher by a small margin than for text-to-foil. This behavior is consistent for both datasets and all five models, with no exceptional anomalies. For non-contrastive explanations, the similarity of NCE-to-fact is consistently higher than NCE-to-foil on the IMDB dataset. However, this is not the case on the medical dataset, as the scores are either equal or slightly higher for NCE-to-foil. For contrastive explanations, the results show that the average CE-to-fact scores on the IMDB dataset are higher than the CE-to-foil scores. This is not as evident in the Medical dataset, as the scores are more equal. Lastly, compared to the baseline (text-to-fact and text-to-foil), the explanations on the IMDB and the Medical dataset are more semantically similar to the text they explain than to the text itself.



(a) IMDB Dataset (b) Medical Dataset

Figure 14: Semantic Similarity BERTScore F1 scores between Fact/Foil label and CE, NCE_{fact}, NCE_{foil} and Text. Scores are presented for all five models.

4.2 Generality

Lexical Diversity Figure 15 and Figure 16 display the average MATTR score for the IMDB and the Medical datasets, respectively. Both figures show the lexical complexity of a text (blue data points) and modified text (orange data points) on the left, while on the right, they display the MATTR scores for CE (blue data points) and MCE (orange data points). First, both CE and MCE are nearly identically diverse lexically on the IMDB dataset (Figure 16). However, there is a more considerable disparity between CE and MCE on the Medical dataset (Figure 16). Moreover, all models have generated similar lexical explanations, observed on the right side of Figure 15, ranging from 0.75 to 0.76 MATTR on average. The range is minimally more extensive on the Medical dataset, up to 0.78, which is the outcome of `Mistral-7B`'s instability compared to the other models. Lastly, there is a more significant gap between the text and the modified text MATTR scores for the Medical dataset than for the IMDB dataset. However, the average scores are much higher on the IMDB dataset than the Medical dataset.

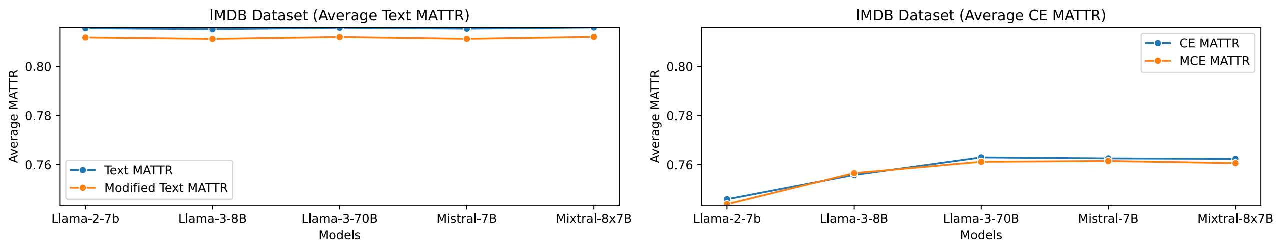


Figure 15: Average MATTR score (window size = 50) for all five models based on the accurate samples for the IMDB dataset. Left: Average MATTR scores for Text and Modified Text. Right: Average MATTR scores for CE and MCE.

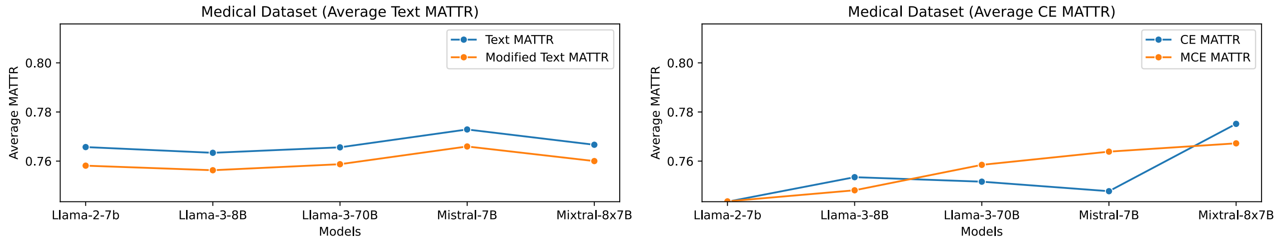


Figure 16: Average MATTR score (window size = 50) for all five models based on the accurate samples for the Medical dataset. Left: Average MATTR scores for Text and Modified Text. Right: Average MATTR scores for CE and MCE.

Lexical Density Figure 17 and Figure 18 represent lexical density (and are very similar to the figures showing lexical diversity), where they show the average density for text (blue) and modified text (orange) on the left side as well as for CE (blue) and MCE (orange) on the right side. Firstly, the lexical density of text and modified text of movie reviews (approximately 0.40 on average for all models) is much lower than the density for medical texts (0.50), as seen on the left side of Figure 17 and Figure 18, respectively. Moreover, there is a more significant disparity between the density of text and modified text in the IMDB dataset than in the Medical dataset. Regarding the explanations, movie review explanations are uniform for all models, ranging from 0.41 to 0.43. In contrast, the medical explanations were less consistent (less dense for `Llama3-70B` with 0.49 on average and more dense for `Mistral-7B` with 0.46 on average). Moreover, the density of explanations is not entirely dependent on the text density. This can be observed by explanations being more dense than text on the IMDB dataset (0.395 vs 0.415); however, explanations are less dense on the medical dataset (0.50 vs 0.48).

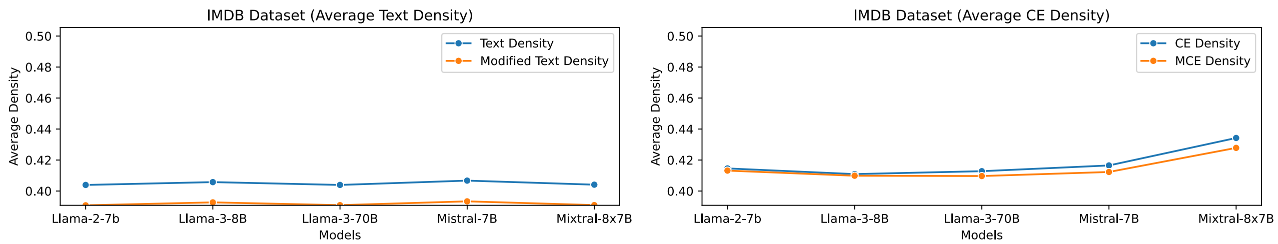


Figure 17: Average density score for all five models based on the accurate samples for the IMDB dataset.

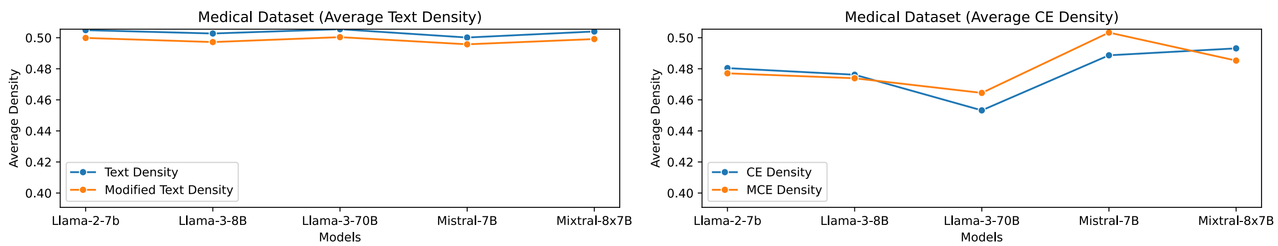


Figure 18: Average density score for all five models based on the accurate samples for the Medical dataset.

Semantic Similarity (Sentence-BERT) Figure 19 presents a comparative analysis of the Sentence-BERT semantic similarity scores (cosine similarity) between the generated contrastive explanations (CE or MCE) (blue) and the original text (text or modified text) (orange) across two datasets: IMDB and Medical. We used violin plots to show the results and visualize numerical data distribution more comprehensively. We noticed that the main drawback of bar plots is that they only show summary statistics, such as means with error bars. They do not reveal the underlying distribution of the data. Violin plots are more effective in displaying data with multiple peaks, which can be obscured in bar plots. The results for each dataset are shown for five different models. Models such as `Llama3-70B` were more consistent in generating similar explanations, as seen by the narrow blue distribution on both datasets. Other models generated samples with a similar distribution as the provided texts. However, there are exceptions. `Llama2-7B` and `Mistral-7B` were less consistent than the text the explanations generated (right side of Figure 19). Lastly, the cosine similarity scores between explanations and the texts are relatively high, roughly 0.75 for the IMDB dataset and 0.90 for the Medical dataset. The medical dataset's cosine similarity of explanations and texts is higher.

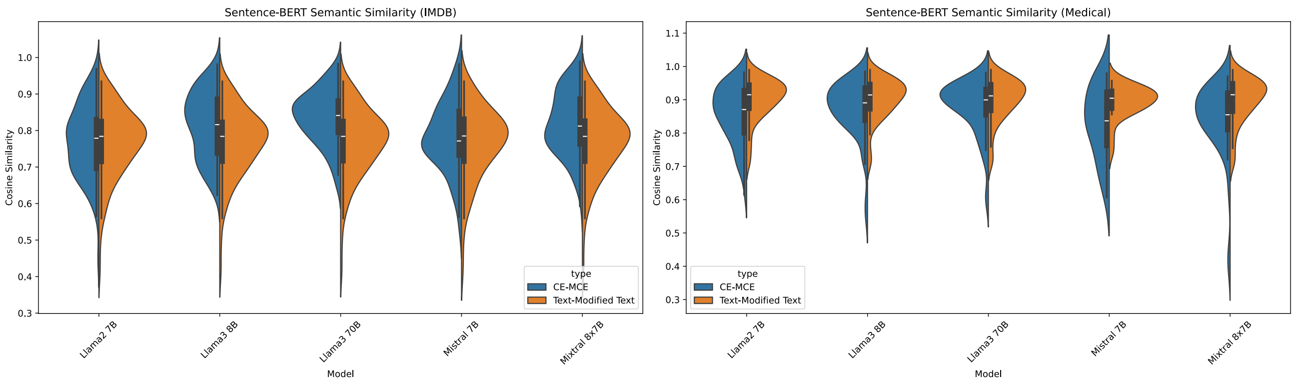


Figure 19: Semantic similarity distributions between two bodies of text: CE-MCE (blue) and Text-Modified Text (orange) using Sentence-BERT cosine similarity. The violin plot on the left represents the IMDB dataset, and the plot on the right represents the Medical dataset.

4.3 Readability

Words per Text Figure 20 and Figure 21 show five sorted plots, each for one model. Each plot shows the lengths of all texts and contrastive and non-contrastive explanations in increasing order, helping to visualize the minimum and maximum lengths. First, the generated explanations (yellow, green and red lines) are generally shorter than the texts based on which those were generated. This is more evident on the IDMB Dataset (Figure 20), with values ranging from 70 words up to 380. The medical dataset does not have such distinction, as the length of explanations varies more. Non-contrastive explanations are the shortest, on average, among all explanations. NCEs explaining foil are shorter than the ones explaining fact. Moreover, contrastive explanations are longer than non-contrastive ones. This is only not the case for `Mistral-7B` on the Medical dataset, as CEs are also shorter on average for `Llama2-7B` on the IMDB dataset.

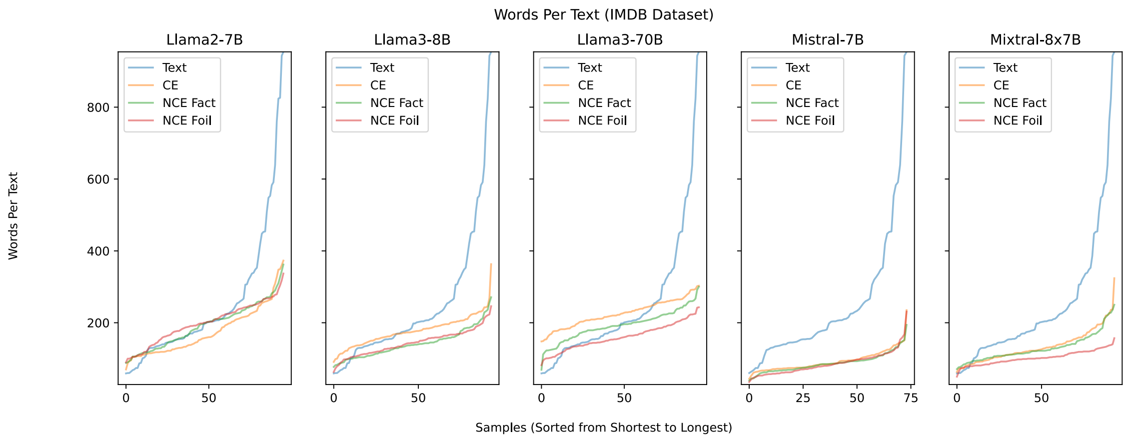


Figure 20: Sorted distribution of samples for Number of Words feature (Readability) in the IMDB dataset for all five models. The samples are sorted in a non-decreasing manner for better visualization.

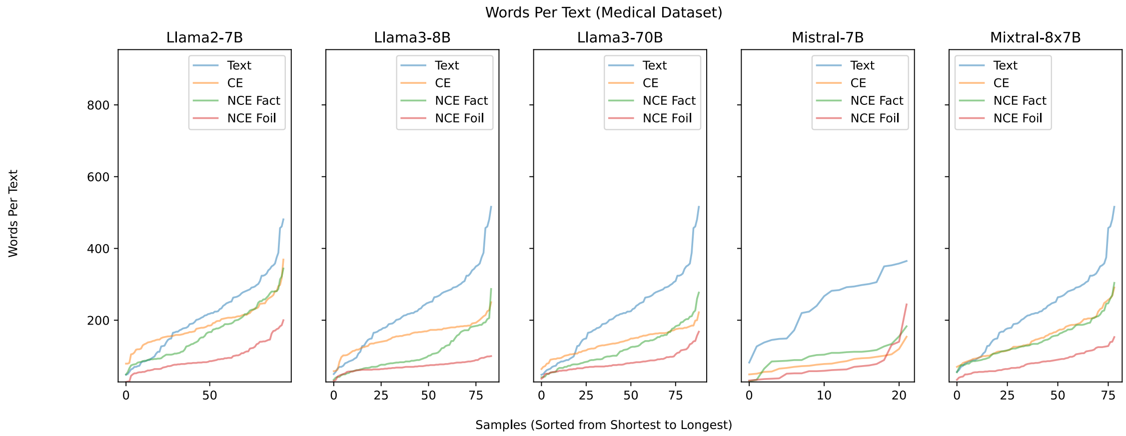


Figure 21: Sorted distribution of samples for Number of Words feature (Readability) in the Medical dataset for all five models. The samples are sorted in a non-decreasing manner for better visualization.

Words per Sentence Figure 22 and Figure 23 are similar to text length; however, these represent the average number of words per sentence. We first observed that the sentence length is consistent across all models for the IMDB and Medical datasets. On average, the lengths do not exceed 40 words per sentence, most of which are roughly 20 words long. There are a few noticeable exceptions. `Mistral-7B` has around ten samples that have sentences of extreme lengths, reaching beyond 140 words per sentence. Lastly, we noticed that generated explanations (yellow, green, and red lines) have sentences slightly longer than the text based on which they are generated (blue line). This is generally the case for all models across both datasets.

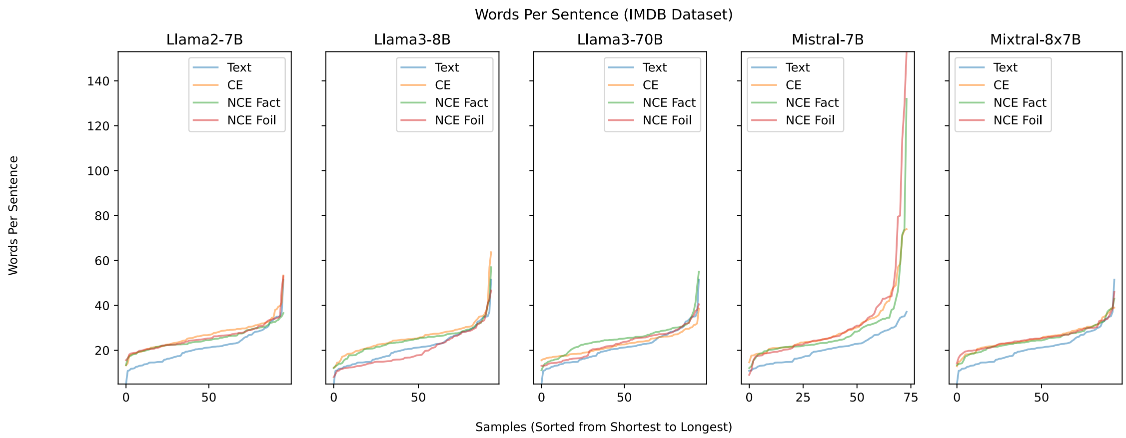


Figure 22: Sorted distribution of samples for Words Per Sentence feature (Readability) in the IMDB dataset for all five models. The samples are sorted in a non-decreasing manner for better visualization.

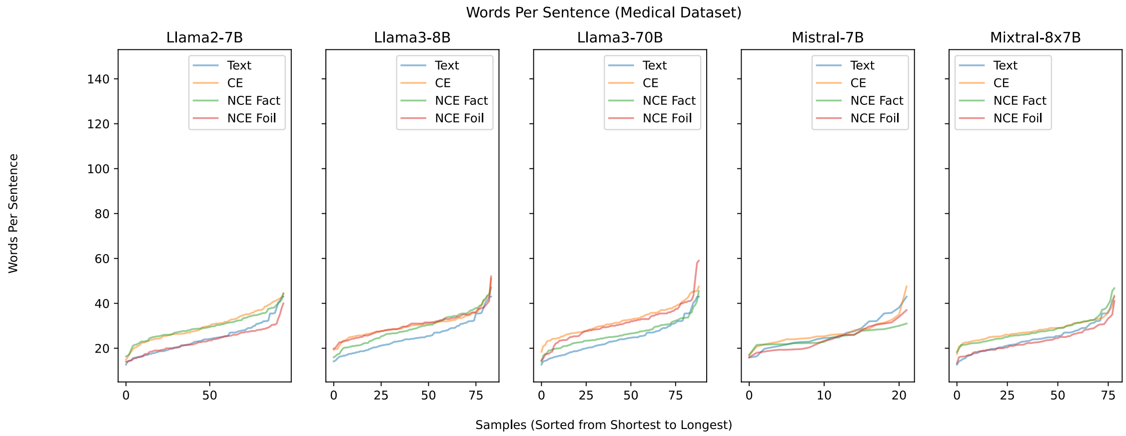


Figure 23: Sorted distribution of samples for Words Per Sentence feature (Readability) in the Medical dataset for all five models. The samples are sorted in a non-decreasing manner for better visualization.

Syllables per Word Figure 22 and Figure 23 represent syllables per word for all samples used for evaluating the models on the IMDB and the Medical dataset, respectively. Moreover, the explanations generally have more syllables per word in the Medical dataset (between 1.3 and 1.9 on average) (Figure 25) than in the IMDB dataset (between 1.3 and 1.5) (Figure 25). This shows that movie review explanations are more consistent in this aspect. Lastly, the average syllables per word ratio is lower in the movie review texts than in the medical diagnoses texts. However, the explanations are not dependent on the texts, as all explanations have higher ratios on the IMDB dataset than the texts, while the opposite is true for the Medical dataset.

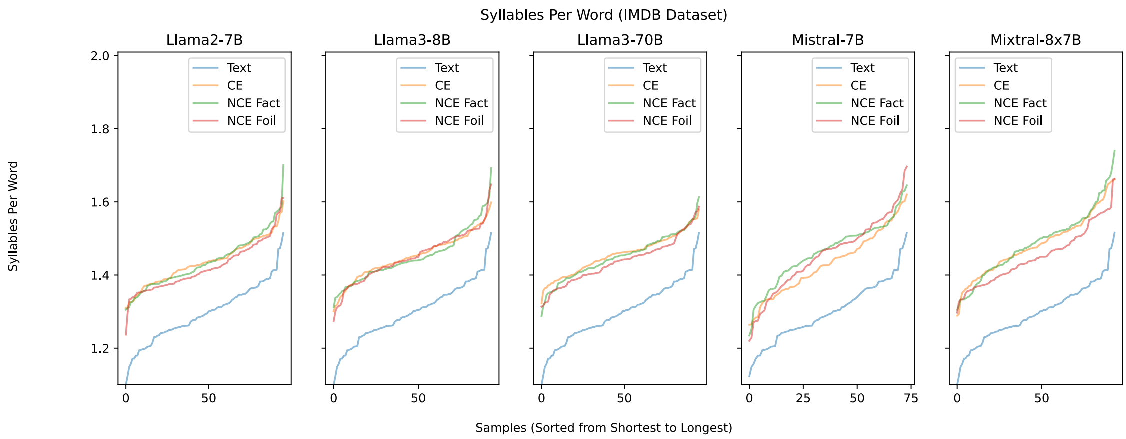


Figure 24: Sorted distribution of samples for Syllables Per Word feature (Readability) in the IMDB dataset for all five models. The samples are sorted in a non-decreasing manner for better visualization.

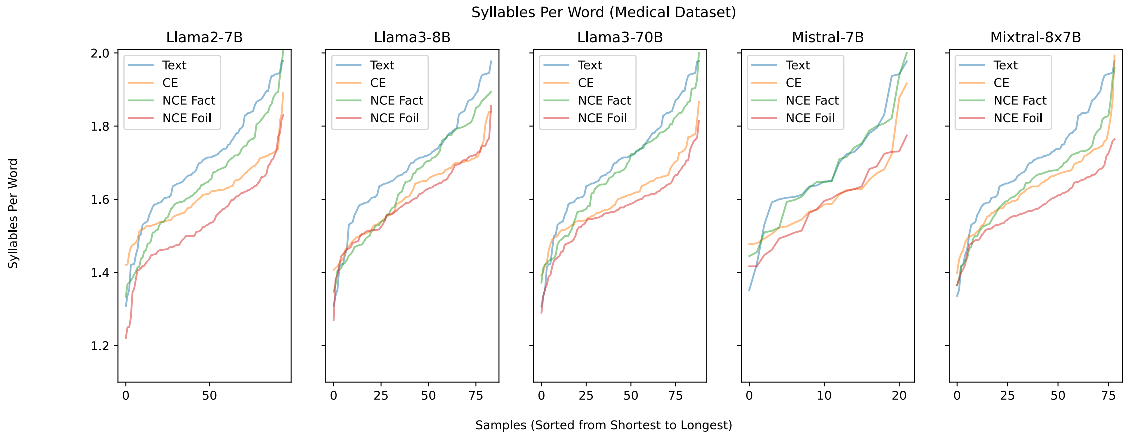


Figure 25: Sorted distribution of samples for Syllables Per Word feature (Readability) in the Medical dataset for all five models. The samples are sorted in a non-decreasing manner for better visualization.

4.4 Questionnaire - The User Study

Figure 26 and Figure 27 show the questionnaire results for two tasks, Movie Reviews and Medical Abstracts, respectively. The results are presented with a stacked bar graph, split into three parts: usefulness, readability, and relevance. The red and yellow colors represent disagreement, while light and dark green colors represent responses that agree with the statements.

Movie Reviews Figure 26 shows generally positive user feedback for Usefulness, Readability and Relevance. All models deemed to be useful in distinguishing between negative and positive sentiments, where at least 70% of the users indicated that (for all models). There were a few disagreements whether the models are useful (2-3 "Strongly disagree"/"Somewhat disagree"). The most readable explanations came from `Llama3-8B`, `Llama3-70B` with 75% responses indicating "Strongly agree". The explanations that came from `Llama3-8B` and `Llama3-70B` were deemed relevant by the users much more often than other models (4 "Somewhat agree" both, and 14 and 13 "Strongly agree", respectively). `Llama2-7B` and `Mistral-7B` both received 16 positive responses, with only 3 "Neutral".

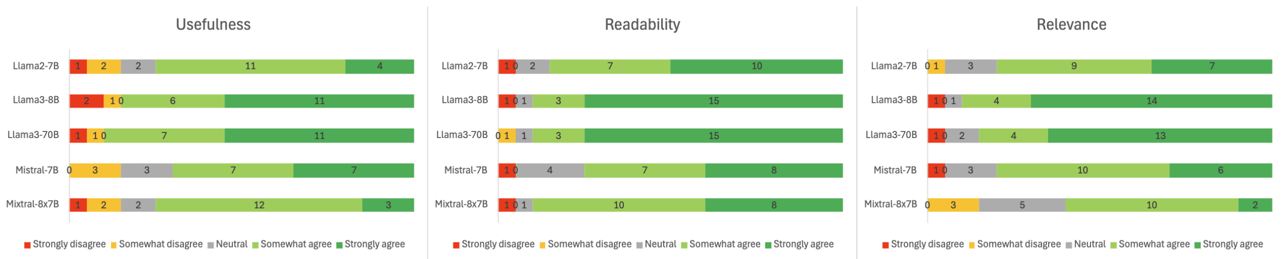


Figure 26: User study results for the Movie Reviews, based on twenty responses. Usefulness (left), Readability (middle), and Relevance (right) are represented with stacked bars. The dark green color on the right represents "Strongly agree" and transitions to "Strongly disagree" on the left, represented by red.

Medical Abstracts According to Figure 27, the results indicate more disagreement with the statement on usefulness, readability and relevance, on average. `Mistral-7B` was considered not useful in giving a contrastive explanation (80% "Strongly disagree", and 20% "Somewhat disagree"). On the other hand, `Llama3-8B` and `Llama2-7B` gave a good contrast, which satisfied 90% of the respondents in both cases. `Llama3-70B` and `Mixtral-8x7` gave mixed responses. Most readable explanations were generated by `Llama3-8B` (only "Strongly agree" or "Somewhat agree"). The majority of participants rated Llama2-7B positively, with 5 responses for "Strongly Agree", 3 for "Somewhat Agree", and 2 "Neutral" responses. `Llama3-70B` had more negative than positive responses, while the `Mistral-7B` and `Mixtral-8x7B` received mixed responses. When it comes to the relevance of the explanation to the given text, `Llama3-8B` had the highest number of positive ratings, with 7 "Somewhat Agree" and 3 "Strongly Agree". `Mixtral-8x7B` also received many positive ratings, with a single expectation for 1 "Strongly disagree". The other models had mixed evaluations, leaning more towards the "Disagree" zone. Moreover, the participants were of the following backgrounds:

- Four BSc
- Two MD or MBBS
- Two MSc (Biomaterials, Tissue Engineering)
- One Resident (Radiology)
- One PhD

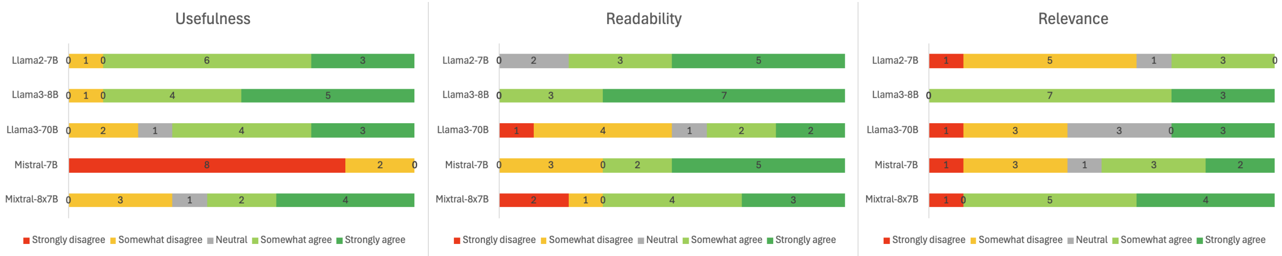


Figure 27: User study results for the Medical Abstracts, based on ten responses. Usefulness (left), Readability (middle), and Relevance (right) are represented with stacked bars. The dark green color on the right represents "Strongly agree" and transitions to "Strongly disagree" on the left, represented by red.

5 Discussion

Our research aimed to investigate how to generate effective contrastive self-explanations for instruction-tuned large language models. We created a method that generates and evaluates contrastive self-explanations called CoSEM (Contrastive Self-Explanations Method). The method consists of five steps: (1) task selection, where a task (dataset) was selected; (2) model accuracy evaluation, where the method tested models on classification/prediction ability; (3) Model self-consistency verification, where the method tested the models for being self-consistent in their responses; (4) Explanations generation, where the method generated contrastive and non-contrastive methods using syntactic modification and model inference, and finally (5) two evaluation approaches. CoSEM split evaluations into automated and user studies (questionnaires). The method evaluates self-explanations on usefulness, generality, readability, and relevance.

Model complexity and dataset affect accuracy and consistency We classified two datasets (IMDB and Medical) to observe how model complexity shapes its accuracy. We noticed a few factors that affected the models' accuracy. First, we have observed that models achieved higher accuracy when predicting labels for movie reviews' sentiment (left side of Figure 13) than when predicting medical diagnoses (right side of Figure 13). We utilized general-use instruction-tuned LLMs, which are not explicitly fine-tuned for specific tasks. The disparity between the accuracy in both tasks can be explained by the vocabulary used in the datasets. The medical diagnoses in the medical dataset utilize vocabulary (words) less commonly used than in movie reviews. This indicates that LLMs have been trained on less medical data than on commonly found movie reviews online (internet users with various backgrounds). Moreover, the IMDB task required the model to predict a sentiment (positive vs negative) rather than a diagnosis (e.g., neoplasms), which is less commonly known by an average person, especially one without a medical background. For such tasks, there exist specifically fine-tuned models, such as Med-PaLM [52], that reach SOTA performance in the medical domain. We believe such models would have been much more accurate for the medical dataset, as they are fine-tuned explicitly to medical vocabulary and context. Secondly, we noticed that model complexity is not a good indicator of its accuracy on datasets. Secondly, model complexity affects the performance in the selected tasks. We have utilized one model with 70B parameters, a few with 7B or 8B parameters, and one 8x7B, which chooses the two best 7B models for inference. We noticed that the 70B parameter model, `Llama3-70B`, performed worse in predicting medical diagnoses than its smaller 7B model predecessor, `Llama2-7B` (see Figure 13). We assumed that more complex models would perform better than smaller models due to exposure to more data and parameters, effectively giving them more power to generate meaningful responses. However, `Llama3-70B` was just as good as `Llama2-7B`, as both achieved 96% on the IMDB dataset (left side of Figure 13) and `Llama3-70B` did worse than `Llama2-7B` on the Medical dataset (right side of Figure 13), where it misclassified six samples more. This behavior could be explained by larger models (70B) potentially overfitting the data they are trained on and encountering issues (falsely classifying data) when generalizing. However, as we only evaluated one LLM with 70B parameters (see section 6), we cannot confirm the same behavior in other 70B parameter models (e.g., `Llama2-70B`). Lastly, model confidence plays a significant role in accurately classifying data. `Mistral-7B` and `Mixtral-8x7B` models achieved lower accuracy due to their lower confidence in prediction. Most of their predictions were inconclusive and therefore marked as falsely classified. `Mistral-7B` only had 22 out of 100 samples on the medical dataset that were correctly classified, much lower than the other models. We have not evaluated the confidence of the responses from the models in this work, but it could be an insightful extension of this method. However, some recent works explore confidence elicitation in LLMs [61]. We believe the model's inability to answer could be preferred over falsely predicting a medical diagnosis, which can have life-threatening consequences, especially in tasks such as medical diagnoses. Finally, we could not perform model self-consistency, described in detail in our limitations (section 6).

LLMs are promising at generating CEs that generalize well We addressed the self-explanations generalization through our generality evaluation, in which we investigated lexical diversity, density, and semantic similarity. First, as we evaluated text diversity with MATTR, we noticed that models generated explanations roughly with the same lexical diversity despite different lexical levels of input text. Figure 15 (right side) showed that all models had nearly identical MATTR scores for CEs and MCEs (blue and orange points overlapping). Similar, although not as perfect, behavior is noticed in medical abstracts (Figure 16), where the scores were less uniform scores (especially Mistral-7B, which was evaluated on fewer samples). The overlap indicates the models' ability to generate CEs that are similarly diverse, regardless of the task. We observed a similar behavior in the lexical density, seen in Figure 17 and Figure 18. The explanations, original and modified, were consistently dense. The explanations were generated around the same level despite lower lexical density in movie reviews (approx. 0.40) and higher in medical abstracts (approx. 0.50). According to Johansson's study [28], the author indicates that adults read the text at a density of around 0.39. Compared to the contrastive explanations achieved through our method, the movie review explanations deviate around this level (0.42). Medical explanations are more lexically dense (0.48), which indicates that not every average adult will find these explanations easily readable. However, we concluded that these results indicate that our self-explanations consistently generalize well lexically. The semantic similarity evaluation indicated larger models (70B) were more robust in generating general contrastive explanations than smaller models. This behavior was observed in Figure 19, where `Llama3-70B` had a much narrower distribution (blue), indicating more consistent results across all samples. This is the case for both tasks and can be addressed by evaluating more samples (see Section 6, Controlling Syntactic Modifications). Finally, we noticed that models were better at generating more general explanations on the medical dataset (Figure 19). This was observed by the higher cosine similarity on the Medical dataset (right side), around 0.90, compared to 0.75 on the IMDB dataset (left side). We observed that the similarity between text and modified text is also much higher in medical explanations. Semantic evaluation, therefore, did not benefit us further in establishing the effectiveness of CEs in their generality. We suspect the modified medical text had fewer modified words (hypernyms and synonyms) than the modified movie text (see Section 6, Model Inference Time).

Contrastive self-explanations are readable We have addressed how easy it would be for users to read contrastive explanations. First, we assumed that contrastive explanations are typically longer than non-contrastive explanations, suggesting they provide more detailed information. Based on Figure 20 and Figure 21, this holds, as on average, CEs (yellow line) had more words than NCEs (green and red lines). Based on Figure 20 and Figure 21, the explanation lengths remained consistently short for both tasks (80-380 words on IMDB and medical datasets), which is beneficial for managing cognitive load during reading. This balance between detail and brevity helps maintain reader engagement without overwhelming them. A few samples are much longer than the rest (see Figure 22 and Figure 23), averaging even at 60 words per sentence. We noticed that some explanations use bullet points, which naturally extend the length of a sentence, as all points are considered single. Using bullet points, for instance, can extend sentence length while making the information accessible and easy to process [40]. This structure helps break down complex information into digestible parts, enhancing readability. However, one must ensure that the bullet points are not too long. Lastly, while more complex texts tend to have a higher syllable count per word, the explanations generated by models do not necessarily mirror this complexity. In Figure 24, we noticed that IDMB texts are roughly at 1.3 syllables per sentence, while medical texts are much higher, at around 1.7 syllables per word (Figure 24). However, movie reviews and medical explanations generalize to around 1.5 syllables per word. This is not necessarily ideal for users who need explanations for movie reviews, as the CEs could use more difficult vocabulary than the texts. On the other hand, it is quite the opposite for the medical texts, as the explanations have less complex vocabulary than the texts.

Evaluating usefulness semantically is challenging We first compared the semantic similarity of texts used to generate the explanations to the labels (fact/foil), seen in the last two groups of columns (right side) 14a and 14b. These scores are supposed to be consistent for the models as we used the same samples (model-independent). However, the difference comes from the difference in accurately predicted samples by the models (so the `Mistral-7B` spike in 14b is most likely caused by a lesser number of text samples used). One might notice that the average similarity of texts to the fact is higher than that of foil, confirming our assumption that texts reflect the fact more than the foil. Nevertheless, we performed statistical analysis as the standard deviation was high. We used the Kolmogorov-Smirnov test with $\alpha = 0.5$, which compared the whole distributions of lists rather than just their means or medians (just like in a paired t-test statistical analysis, which we have not used due to the distributions of scores not being normally distributed). The tests indicated no difference between text-fact and text-foil similarity for the IMDB dataset. On the Medical dataset, only the texts used for `Mistral-7B` and `Mistral-8x7B` indicated that those are more semantically similar to the fact label. However, the small amount of samples makes the results less trustworthy than those with more samples. Therefore, we have realized that texts are semantically as similar to the foil as they are to the fact label We also evaluated non-contrastive explanations, which only should answer the fact or foil labels exclusively. The middle columns of 14a showed that NCE is more semantically similar to the foil, with an exception for `Llama3-70B` model. which shows no significant difference. After closer inspection of evaluated samples, we noticed that models use similar arguments when giving a non-contrastive explanation. For example, one NCE_{fact} explanation stars with *"This text is negative for several reasons:[...]"* while NCE_{foil} starts with *"This text is not positive for several reasons:[...]"*. Both continue using a similar explanation approach, explaining the fact and giving more semantic similarity to it. 14b for the Medical dataset showed that statistically, the NCEs are as similar to the fact as they are to the foil. Overall, we conclude that the F1 scores for both datasets are nearly identical and that, in practice, both explanations for the medical and IMDB tasks are equally similar to both labels. We need to deploy semantic similarity evaluations further to determine NCEs' nature. We also checked how the contrastive self-explanations perform with semantic similarity to both labels. We hypothesized that the semantic similarity of CE-to-fact and CE-to-foil is approximately equal. The results indicate that CE's similarity to the fact on the IMDB dataset is more significant than the one on the foil. Although the difference is significant, the similarity scores are still relatively close to each other, with scores ranging from 0.28 to 0.34 for fact and 0.26-0.315 for foil, technically indicating more similarity for the fact (see Figure 14a, the first two column groups). In Figure 14b, the CEs are also semantically closer to the fact label. However, the difference is statistically irrelevant (except for the undersampled `Mixtral-8x7B` model, which shows that statistically, there is a difference). Therefore, based on the results, we concluded that the contrastive explanations are as semantically similar to the fact as they are to the foil, indicating that the contrast is similar. This does not provide any meaningful insights on what explanation parts show the contrast.

Users find contrastive self-explanations mostly readable, useful, and relevant Some models were quite good at generating contrastive self-explanations that are useful (contrastive), readable, and relevant, while others' self-explanations are not as effective.

For the IMDB dataset Figure 26, `Llama3-8B` and `Llama3-70B` consistently received the most positive feedback across all three categories, especially in readability and relevance, where most participants rated them as "Strongly Agree". `Mistral-7B` generally received mixed reviews, with a combination of positive and neutral feedback and some negative ratings, particularly for relevance and usefulness. `Mixtral-8x7B` and `Llama2-7B` showed a mix of positive and negative responses, especially for usefulness and relevance, but were generally well-rated for readability. Overall, `Llama3-8B` and `Llama3-70B` stood out as the top performers in this dataset, while `Mistral-7B` showed a more divided reception.

For the Medical dataset Figure 27, `Llama3-8B` consistently received the most positive ratings across all three categories (Usefulness, Readability, Relevance), suggesting it was the best performing model overall. `Mistral-7B` has the most negative feedback, particularly for usefulness (8 Strongly Disagree responses), though its readability and relevance ratings are more mixed. `Mixtral-8x7B` and `Llama2-7B` had more mixed responses, with both positive and negative feedback, but generally leaned towards positive ratings, especially for readability. `Llama3-70B` showed a balance of positive and negative feedback across the categories, with more neutral or slightly negative responses for readability and relevance.

Other observations We have further made the following observations. First, general-purpose instruction-tuned LLMs are not specifically tailored to specific tasks (e.g., medical). We believe that fine-tuned task-specific models for such domains could further enhance the effectiveness of contrastive self-explanation. However, general use cases that do not require domain expertise (e.g., movie reviews) and SOTA models, such as those from the Llama family, perform well. Moreover, we cannot guarantee that such fine-tuned models would provide more effective (i.e., contrastive/useful, general, readable, intelligible, and relative) contrastive explanations. A further study would need to be performed to investigate that. The authors of Mistral 7B, fine-tuned to follow instructions, claimed that the model surpasses the Llama 2 13B on various benchmarks [27]. Although this work utilizes the 7B model and not the 13B model, the results show otherwise. The Llama2 7B model performed better in model accuracy when asked for production on two different tasks and was more confident in its responses. Despite the accuracy hiccup, the Mistral 7B model performs well, but in contrastive self-explanations, it does not outperform the Llama models.

Summary By addressing the above points, we have evaluated the effectiveness of contrastive self-explanations generated by general-use instruction-tuned large language models. We have discussed how the model complexity and dataset affect accuracy and consistency. We have also analyzed how promising LLMs are at generating CEs that generalize well. We showed that CEs are readable to users, although the task for which those are generated affects the readability. However, evaluating usefulness automatically through semantic similarity is challenging. Nevertheless, our questionnaire showed that humans see the explanations to contrast fact and foil. Moreover, users deem some LLMs to be good explainers, as they provide useful (contrastive), readable, and relevant contrastive self-explanations. Our proposed method generates effective self-explanations and evaluates various features that give insight into how LLMs can be utilized for contrastive explanations in multiple tasks. We have evaluated the explanations in an automated manner and through a human study, which, as far as we know, was done in a computerized manner without a human evaluation. This has helped us ensure that the models consistently output contrastive self-explanation and that those explanations are helpful to the users.

6 Limitations and Future Work

This section outlines the primary limitations encountered in our research and proposes directions for future work to enhance the generation and application of contrastive explanations using LLMs. Despite achieving significant insights into the efficacy of current models, several constraints limited our exploration and shaped our results. Additionally, we explore the potential direction to enhance and broaden the application of contrastive explanations.

Questionnaire Budget Utilizing crowd-source workers seems relatively straightforward. However, a budget is required to guarantee a certain level of quality in the respondents' answers. Unfortunately, last-minute decision has been made to abandon a crowd-sourced approach and conduct a small-scale user study. Initially, we aimed to utilize the Prolific platform to gather responses. The recommended pay is "at least £9/$12 per hour"[8] to ensure meeting ethical standards and data quality. The estimated cost of paying for 100 respondents, given that each sample has one text and five explanations, is roughly €300. In the end, we collected only twenty responses for the movie reviews and ten responses fo the medical abstracts, which required participants to be in the medical field. Given a budget, crowd-sourced workers with medical qualifications could have been employed for this questionnaire, resulting in more responses (possibly giving insights into how their fields affect their interpretation of contrastive self-explanations. The lack of budget implied limiting the number of samples we ran. Ideally, we would run the experiment on all available samples correctly classified by the models.

Model Inference Time The most time-consuming part was running the model accurately and generating all self-explanations for the original and modified text. Initially, we intended to utilize six models: two Llama2 (7B and 70B variants), two Llama3 (8B and 70B), one Mistral (7B) and one Mixtral (8x7B). However, we ended up eliminating Llama2-70B due to memory issues encountered during the inference. On average, `Llama2-70B` and `Llama3-70B` took roughly 1 hour and 30 minutes to generate all self-explanation for a single sample for a single dataset. However, Llama2 would through an out-of-memory error after running a single sample. Due to this behavior and the lack of access to more GPU computer power, we have removed this model from our evaluation. Secondly, the inference time, especially for the 70B models, took much longer than for smaller models (7B and 8B). The smaller models, on average, took 3 hours to generate all 100 samples in a dataset. That is significantly less time than for the large models. We had limited access to the supercomputer on which the inference took place, so we collected only 100 samples for each model. Given more time and resources, the evaluation would have constituted more samples, ideally all available in the datasets.

Controlling Syntactic Modifications While evaluating the generality of contrastive explanations, we noticed a more considerable semantic similarity between the modified and contrastive explanations in the medical dataset. After further inspection, we noticed that the modified text in the medical dataset had fewer substituted words (hypernyms and synonyms) than the one in the IMDB dataset. Therefore, we should have tackled this fairness issue by adding control mechanisms that would always modify a certain percentage of words in a given text. This would result in a fair evaluation of generality in various tasks.

Self-Consistency Evaluation Issues In our CoSEM, we have adapted the self-consistency method from [39] and tried adapting it for contrastive self-explanations to ensure that the explanations provided are valid (see Figure 6). The authors of that paper have explored three approaches to checking for self-consistency (see Section 2). We have adapted the method to use redaction, which, in theory, should have helped us force the model to mask tokens required to generate a contrastive self-explanation. In practice, the models would output completely modified texts that do not reflect the original text and hence do not serve in further classification with redacted tokens. Sometimes, the models would generate non-sense responses such as "Why [REDACTED] and not [REDACTED]?" instead of a text with some tokens masked. However, the authors of the method also noticed that the model self-explanation should not be generally trusted. We have decided not to include self-consistency in our method for two reasons. First, it did not directly affect the generated contrastive explanations. Second, it required significant time during inference, which tied in with the previous limitation. To quantitatively evaluate self-consistency, we had to use different model seeds. Each iteration involved loading the model once, running inference once, and running classification once. That would have added a big-time overhead to the inference and our experiment, which we did not have.

[8] https://www.prolific.com/resources/how-much-should-you-pay-research-participants

Explanations from Incorrect Models In this study, we disregarded evaluating contrastive self-explanation if the model failed to classify the contrast label to a given text correctly. In this study, we have assumed that a model is not entirely "trustworthy" if it does not indicate the right label for the text it explains. However, such a model can still answer a contrastive question, which could partially contribute to explaining a situation (text). One might argue that even such an explanation might be compelling in exploring the contrast, but that would further require evaluating how much trust a user puts into such an explanation. But for now, the explanations do not precisely align with the reasoning processes of the LLMs [2]. The topic of trustworthiness in LLMs is an exciting direction that can enhance CoSEM.

CE-tuned Models A potential direction to consider is the development of models specifically fine-tuned to generate contrastive explanations by default. This approach would involve fine-tuning chat models by using contrast (fact and foil) in the responses. Such LLMs could offer more insightful contrasts in their outputs, thereby improving decision-making support in various tasks. (e.g., medicine, law). The potential for deploying these CE-tuned models across various domains shows a promising path for expanding the effectiveness and applicability of LLM contrastive self-explanations in real-world scenarios.

Model Choice and CE Real-World Applications Besides fine-tuning models for CE, utilizing task-tuned models rather than general-use ones could result in even more effective contrastive self-explanations. Models such as Med-PaLM [52] created by Google outperform general-use models. Combined with CE-tuning, such models could be a promising companion for medical staff in finding explanations for medical diagnoses (such as differential diagnoses).

7 Conclusion

This research introduced "CoSEM: Contrastive Self-Explanation Method" for instruction-tuned large language models. The method generates contrastive self-explanations and evaluates them in an automated manner and through questionnaires. In this work, we propose a novel, non-gradient-based method for generating contrastive self-explanations and evaluating them automatically and through user feedback. The method does not require access to the model internals, which is ideal for black-box models. We have adapted an existing self-consistency method to a contrastive setting for self-explanations. We created a method to syntactically modify text in a rule-based fashion, which allows for evaluating contrastive self-explanations on their generality. Furthermore, we introduce a model inference method that generates contrastive and non-contrastive self-explanations. Finally, we evaluate how contrastive, general and readable contrastive self-explanations are and how intelligible, useful and relevant users find them.

Our study explored the generation of effective contrastive explanations using LLMs, focusing on the complexities of various datasets and model capabilities. Through our comprehensive analysis across different domains — IMDB for general sentiment and a specialized medical dataset — we have established vital findings that enhance our understanding of LLM behavior in generating contrastive self-explanations. We have evaluated contrastive self-explanations generated by large language models on their usefulness, generality, readability, intelligibility, and relevance.

First, we found out that model complexity impacts prediction accuracy. Models with more parameters do not guarantee better performance, as this also depends on the task we the models for. Models find it easier to predict correctly more general tasks, such as identifying the sentiment of movie reviews, than particular tasks, such as predicting medical diagnoses. This does not necessarily involve model complexity but how "chat LLMs" are tuned to the data. Using task-specific tasks could improve performance on specific tasks (i.e., medical diagnoses), but it is uncertain if such a model would perform as well on more general knowledge tasks. Nevertheless, ensuring high model accuracy before using it for life-threatening tasks should be a priority.

Second, we found that models are good at generalizing contrastive self-explanations. We have achieved this by syntactically modifying text samples to generate these explanations. The explanations generalize content well across all models and all tasks we used. All criteria we used for generality (lexical diversity, lexical density, and semantic similarity) indicated that the explanations are similarly diverse and dense lexically across all models, where density was more correlated with the text used to generate the explanations. Overall, this indicates that all models are well suited to generalizing explanations of similar complexity to the text they are given. Larger (more complex) models were marginally more consistent in generalizing the explanations than smaller models.

Furthermore, our readability evaluation indicated that contrastive self-explanations are desirable for the reader due to their consistently short length (compared to the text used to generate them). Moreover, the explanations remained consistently short for both tasks, which indicates their ability not to stretch the cognitive load while reading them. However, the text readability might change based on the selected task, as the readability scores indicated that explanations were higher than the text for a more general task. This, however, was not the case for a more specific task, as the explanations were more accessible to read than the text.

Generated contrastive self-explanations have a good balance of contrast in them. Our results show that LLMs generate CE that has a balance of both the fact and foil labels. There was no statistically significant preference in the explanations toward either task label. However, models are not as good at generating non-contrastive explanations as those generally deviate towards one label, even if asked to explain using the other. Asking for an explanation that uses foil would boil down to an explanation that uses fact with a negation, indicating the same content in the central part of the explanation.

The user study has revealed that contrastive self-explanations generated by different models indicate significant variations in their effectiveness across both the IMDB and Medical datasets. Models like `Llama3-8B` and `Llama3-70B` consistently stand out, receiving the highest usefulness, readability, and relevance ratings. These models were particularly strong in readability and relevance, where most users strongly agreed with their quality. On the other hand, `Mistral-7B` received mixed to negative feedback, especially for usefulness, indicating its self-explanations were less impactful. `Mixtral-8x7B` and `Llama2-7B` displayed more diverse results, with positive and negative responses across the categories, though they performed relatively well in readability. `Llama3-8B` is the top-performing model across both datasets, showcasing its ability to generate well-rounded and effective self-explanations. At the same time, `Mistral-7B` exhibits a more polarized reception, pointing to potential improvements in its usefulness and relevance.

Although our results are promising, we have also encountered some limitations. Conducting a comprehensive, rigorous, and robust user study requires resources such as a financial budget, time, and participants. We have encountered a budget limitation requiring us to reduce the number of crow-source participants. Similarly, finding suitable candidates for tasks that require a specific background (i.e., medical dataset) is difficult under time constraints. Second, running model inference is a task that requires access to resources. Despite having access to a supercomputer with a GPU that runs such inference, we could only run a limited number of samples. Third, we encountered an issue with self-consistency in models, similar to [39]. Self-consistency in model responses requires further research and is a separate topic that needs additional attention.

Despite the limitations of questionnaire budget, model inference time, self-consistency, and incorrect model explanations, we see possible future directions for researching LLM-generated contrastive self-explanations. Finding the budget for a more significant user study and resources to run inference on more samples should result in a more extensive CoSEM evaluation. Fine-tuning models to generate contrastive self-explanations is a potential path for giving users even more useful (contrastive) explanations. Such models could be more fruitful for answering user questions by providing a contrast to help them understand specific tasks. Finally, introducing CoSEM to a broader set of tasks can become tremendously helpful in domains that require explanations, such as medicine or law.

Our study contributes to the literature by dissecting the mechanisms behind effective contrastive explanation generation in LLMs and suggesting pathways for future research, particularly in integrating domain-specific knowledge and model confidence. The groundwork laid by this research promises to guide future explorations into enhancing LLM utility across diverse applications.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. *arXiv preprint arXiv:2402.04614*, 2024.

[3] I Elaine Allen and Christopher A Seaman. Likert scales and data analyses. *Quality progress*, 40(7):64–65, 2007.

[4] Gilles Audemard, Jean-Marie Lagniez, Pierre Marquis, and Nicolas Szczepanski. On contrastive explanations for tree-based classifiers. In *ECAI 2023*, pages 117–124. IOS Press, 2023.

[5] Henry Bae, Aghyad Deeb, Alex Fleury, and Kehang Zhu. Complexitynet: Increasing llm inference efficiency by learning task complexity. *arXiv preprint arXiv:2312.11511*, 2023.

[6] Jawid Ahmad Baktash and Mursal Dawodi. Gpt-4: A review on advancements and opportunities in natural language processing. *arXiv preprint arXiv:2305.03195*, 2023.

[7] Yves Bestgen. Measuring lexical diversity in texts: The twofold length problem. *Language Learning*, 2023.

[8] Lorenzo Betti, Carlo Abrate, Francesco Bonchi, and Andreas Kaltenbrunner. Relevance-based infilling for natural language counterfactuals. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 88–98, New York, NY, USA, 2023. Association for Computing Machinery.

[9] Saneem Chemmengath, Amar Prakash Azad, Ronny Luss, and Amit Dhurandhar. Let the cat out of the bag: Contrastive attributed explanations for text. *arXiv preprint arXiv:2109.07983*, 2021.

[10] Liang Chen, Pei Xu, and De Liu. Experts versus the crowd: a comparison of selection mechanisms in crowdsourcing contests. *SSRN Electronic J.(epub)*, 2015.

[11] Courtney D Corley and Rada Mihalcea. Measuring the semantic similarity of texts. In *Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment*, pages 13–18, 2005.

[12] Michael A Covington and Joe D McFall. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100, 2010.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[14] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives, 2018.

[15] Christiane Fellbaum. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer, 2010.

[16] Rudolph Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221, 1948.

[17] Alan Garfinkel. Forms of explanation: Rethinking the questions in social theory. 1982.

[18] Stanford NLP Group et al. Stanza–a python nlp package for many human languages, 2018.

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

[20] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021.

[21] Steffen Herbold. Semantic similarity prediction is better than other semantic similarity measures. *arXiv preprint arXiv:2309.12697*, 2023.

[22] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.

[23] Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H Gilpin. Can large language models explain themselves? a study of llm-generated self-explanations. *arXiv preprint arXiv:2310.11207*, 2023.

[24] Emmanuel M Ikart. Survey questionnaire survey pretesting method: An evaluation of survey questionnaire via expert reviews technique. *Asian Journal of Social Science Studies*, 4(2):1, 2019.

[25] Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. Contrastive explanations for model interpretability. *arXiv preprint arXiv:2103.01378*, 2021.

[26] Andrew T Jebb, Vincent Ng, and Louis Tay. A review of key likert scale development advances: 1995–2019. *Frontiers in psychology*, 12:637547, 2021.

[27] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[28] Victoria Johansson. Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working papers/Lund University, Department of Linguistics and Phonetics*, 53:61–79, 2008.

[29] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5):1–29, 2022.

[30] James C Kaufman, John Baer, Jason C Cole, and Janel D Sexton∗. A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal*, 20(2):171–178, 2008.

[31] Ruth Kircher and L Zipp. Questionnaires to elicit quantitative data. *Research methods in language attitudes*, pages 129–144, 2022.

[32] Franklin Mark Liang. *Word Hy-phen-a-tion by Com-put-er*. Citeseer, 1983.

[33] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[34] Peter Lipton. Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27:247–266, 1990.

[35] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[36] Ronny Luss, Erik Miehling, and Amit Dhurandhar. Cell your model: Contrastive explanation methods for large language models. *arXiv preprint arXiv:2406.11785*, 2024.

[37] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[38] Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan Saha. Generate your counterfactuals: Towards controlled counterfactual generation for text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13516–13524, 2021.

[39] Andreas Madsen, Sarath Chandar, and Siva Reddy. Can large language models explain themselves? *arXiv preprint arXiv:2401.07927*, 2024.

[40] Robert Michielutte, Judy Bahnson, Mark B Dignan, and Elissa M Schroeder. The use of illustrations and narrative text style to improve readability of a health education brochure. *Journal of Cancer Education*, 7(3):251–260, 1992.

[41] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[42] Tim Miller. Contrastive explanation: A structural-model approach. *The Knowledge Engineering Review*, 36:e14, 2021.

[43] Raghuvar Nadig, J Ramanand, and Pushpak Bhattacharyya. Automatic evaluation of wordnet synonyms and hypernyms. In *Proceedings of ICON-2008: 6th International Conference on Natural Language Processing*, volume 831. Citeseer, 2008.

[44] Letitia Parcalabescu and Anette Frank. On measuring faithfulness of natural language explanations. *arXiv preprint arXiv:2311.07466*, 2023.

[45] Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. Is temperature the creativity parameter of large language models? *arXiv preprint arXiv:2405.00492*, 2024.

[46] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[47] Donya Rooein, Paul Rottger, Anastassia Shaitarova, and Dirk Hovy. Beyond flesch-kincaid: Prompt-based metrics improve difficulty classification of educational texts. *arXiv preprint arXiv:2405.09482*, 2024.

[48] Alexis Ross, Ana Marasović, and Matthew E Peters. Explaining nlp models via minimal contrastive editing (mice). *arXiv preprint arXiv:2012.13985*, 2020.

[49] Lucas Shen Y. S. LSYS/LexicalRichness: v0.5.1, August 2023.

[50] Tim Schopf, Daniel Braun, and Florian Matthes. Evaluating unsupervised text classification: Zero-shot and similarity-based approaches. In *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '22, page 6–15, New York, NY, USA, 2023. Association for Computing Machinery.

[51] Hunny Sharma. How short or long should be a questionnaire for any research? researchers dilemma in deciding the appropriate questionnaire length. *Saudi journal of anaesthesia*, 16(1):65–68, 2022.

[52] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[53] John Sweller. Cognitive load theory. In *Psychology of learning and motivation*, volume 55, pages 37–76. Elsevier, 2011.

[54] Teerapaun Tanprasert and David Kauchak. Flesch-kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14, 2021.

[55] Sule Tekkesinoglu and Lars Kunze. From feature importance to natural language explanations using llms with rag. *arXiv preprint arXiv:2407.20990*, 2024.

[56] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[57] Eric WK Tsang and Florian Ellsaesser. How contrastive explanation facilitates theory building. *Academy of Management Review*, 36(2):404–419, 2011.

[58] Bas C Van Fraassen. *The scientific image*. Oxford University Press, 1980.

[59] Jason Weston and Sainbayar Sukhbaatar. System 2 attention (is something you might need too). *arXiv preprint arXiv:2311.11829*, 2023.

[60] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. *arXiv preprint arXiv:2101.00288*, 2021.

[61] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

[62] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.

[63] Kayo Yin and Graham Neubig. Interpreting language models with contrastive explanations. *arXiv preprint arXiv:2202.10419*, 2022.

[64] Fred Zenker and Kristopher Kyle. Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, 47:100505, 2021.

[65] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

A Questionnaire

The questionnaire presents five contrastive explanations from different LLMs. The order of the models is the following:

1. Explanation 1 `meta-llama/Meta-Llama-3-8B-Instruct`
2. Explanation 2 `meta-llama/Meta-Llama-3-70B-Instruct`
3. Explanation 3 `meta-llama/Llama-2-7b-chat-hf`
4. Explanation 4 `mistralai/Mistral-7B-Instruct-v0.1`
5. Explanation 5 `mistralai/Mixtral-8x7B-Instruct-v0.1`

A.1 Questionnaire Template

This is a generic template format used for questionnaires for both the medical and movie review datasets. The scenario differs for each dataset, wherein one LLM is an AI system for predicting diagnoses, and in another, it predicts the sentiment of a movie review.

Scenario

[Scenario]

Text The AI systems are given the following information (text):

[Text]

Based on the text above, you are given 5 explanations provided by different AI systems. Answer all 3 questions provided with each explanation

Explanation 1

[CE]

1. The explanation is **useful** in distinguishing between [fact label] and [foil label].

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

2. The explanation is **readable** to me (I can read it easily with comprehension).

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

3. The explanation provides **relevant** information (based on the text).

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

Explanation 2

[CE]

1. The explanation is **useful** in distinguishing between [fact label] and [foil label].

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

2. The explanation is **readable** to me (I can read it easily with comprehension).

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

3. The explanation provides **relevant** information (based on the text).

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

Explanation 3

[CE]

1. The explanation is **useful** in distinguishing between [fact label] and [foil label].

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

2. The explanation is **readable** to me (I can read it easily with comprehension).

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

3. The explanation provides **relevant** information (based on the text).

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

Explanation 4

[CE]

1. The explanation is **useful** in distinguishing between [fact label] and [foil label].

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

2. The explanation is **readable** to me (I can read it easily with comprehension).

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

3. The explanation provides **relevant** information (based on the text).

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

Explanation 5

[CE]

1. The explanation is **useful** in distinguishing between [fact label] and [foil label].

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

2. The explanation is **readable** to me (I can read it easily with comprehension).

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

3. The explanation provides **relevant** information (based on the text).

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

A.2 Medical Dataset - Concrete Example

Scenario

A hospital is testing a new AI system that is supposed to explain to a doctor what medical condition the patient has from a text excerpt. It does this by a so-called contrastive explanation. Those explanations aim to explain why a specific condition was chosen over another. The system outputs five different explanations, which you must evaluate as a doctor with medical knowledge.

Text The AI systems are given the following information (text):

Coronary trapping of a complement activation product (C3a des-Arg) during myocardial reperfusion in open-heart surgery. Accumulation of complement factors has been found to occur in the myocardium after infarction. We studied the possibility that the complement activation product C3a des-Arg is trapped within the coronary circulation during reperfusion of the ischemic myocardium. In 11 patients undergoing routine coronary artery bypass grafting, arterial blood was sampled before, during and after cardiopulmonary bypass. Blood was drawn from the coronary sinus concomitantly with arterial blood sampling 5 and 30 min after release of the aortic cross-clamp (n = 10). From a preoperative value of 92 +/- 13 ng/ml, C3a des-Arg rose during CPB to a maximum of 1816 +/- 393 at the end of CPB. Following reperfusion for 5 min, C3a des-Arg was 1284 +/- 232 ng/ml in arterial and 1106 +/- 100 in coronary sinus blood, a significant difference (p less than 0.05). The amount of C3a des-Arg trapped in the heart at 5-min reperfusion showed positive correlation with its arterial concentration (p less than 0.05). No significant difference was found after 30 min of reperfusion. Complement activation products trapped in the heart in the early reperfusion period may play a pathogenetic role in myocardial ischemia-reperfusion injury.

Based on the text above, you are given 5 explanations provided by different AI systems. Answer all 3 questions provided with each explanation.

Explanation 1

The text you provided is discussing the accumulation of complement factors in the myocardium after infarction, specifically during open-heart surgery. The focus of the study is on the trapping of complement activation product C3a des-Arg within the coronary circulation during reperfusion of the ischemic myocardium.
The term "cardiovascular diseases" refers to a broad category of conditions that affect the cardiovascular system, including heart disease, stroke, and vascular disease. In this context, "cardiovascular diseases" would encompass a wide range of conditions, including myocardial infarction (heart attack), which is the focus of the study you provided.
On the other hand, "neoplasms" refers to abnormal growths or tumors that can occur in any part of the body. While neoplasms can affect the cardiovascular system, they are not the primary focus of the study you provided, which is specifically focused on myocardial ischemia-reperfusion injury.
Therefore, the term "cardiovascular diseases" is more appropriate in this context, as it encompasses the specific condition being studied (myocardial infarction) as well as other related conditions that affect the cardiovascular system.

1. The explanation is **useful** in distinguishing between *general pathological conditions* and *neoplasms*.

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

2. The explanation is **readable** to me (I can read it easily with comprehension).

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

3. The explanation provides **relevant** information (based on the text).

1	2	3	4	5
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

Explanation 2

The text is focused on cardiovascular diseases, specifically myocardial ischemia-reperfusion injury, because the study is investigating the role of complement activation products in the development of this condition. Myocardial ischemia-reperfusion injury is a type of cardiovascular disease that occurs when the heart muscle is deprived of oxygen (ischemia) and then reperfused with oxygenated blood, leading to damage and inflammation.
The study is examining the accumulation of complement factors in the myocardium after infarction and the trapping of these factors in the coronary circulation during reperfusion. The researchers are investigating the potential role of these factors in the development of myocardial ischemia-reperfusion injury, which is a cardiovascular disease.
Neoplasms, on the other hand, are abnormal growths of cells that can be benign or malignant. While neoplasms can also be associated with cardiovascular disease, the focus of this study is on the cardiovascular disease itself, rather than the neoplasms.

1. The explanation is **useful** in distinguishing between *general pathological conditions* and *neoplasms*.

1	2	3	4	5
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

2. The explanation is **readable** to me (I can read it easily with comprehension).

1	2	3	4	5
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

3. The explanation provides **relevant** information (based on the text).

1	2	3	4	5
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

Explanation 3

The reason it is cardiovascular diseases and not neoplasms is because the text is discussing a study related to heart surgery (coronary artery bypass grafting) and the effects of reperfusion on the heart, specifically the trapping of a complement activation product (C3a des-Arg) in the coronary circulation. The study is investigating the potential role of this product in myocardial ischemia-reperfusion injury, which is a type of cardiovascular disease.
Neoplasms, on the other hand, refer to abnormal growths of tissue, such as tumors, which are not related to the topic of this study. The study is focused on the cardiovascular system and the effects of reperfusion on the heart, rather than on cancer or tumor growth.

1. The explanation is **useful** in distinguishing between *general pathological conditions* and *neoplasms*.

1	2	3	4	5
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

2. The explanation is **readable** to me (I can read it easily with comprehension).

1	2	3	4	5
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

3. The explanation provides **relevant** information (based on the text).

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

Explanation 4

The text is discussing the accumulation of complement factors in the myocardium after infarction and the possibility that the complement activation product C3a des-Arg is trapped within the coronary circulation during reperfusion of the ischemic myocardium. The text is focused on cardiovascular diseases, specifically myocardial ischemia-reperfusion injury, and not on neoplasms.

1. The explanation is **useful** in distinguishing between *general pathological conditions* and *neoplasms*.

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

2. The explanation is **readable** to me (I can read it easily with comprehension).

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

3. The explanation provides **relevant** information (based on the text).

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

Explanation 5

The text you provided is discussing a study on coronary trapping of a complement activation product (C3a des-Arg) during myocardial reperfusion in open-heart surgery. The researchers are interested in studying the possibility that C3a des-Arg is trapped within the coronary circulation during reperfusion of the ischemic myocardium. The reason this study focuses on cardiovascular diseases (specifically myocardial ischemia-reperfusion injury) rather than neoplasms is because of the nature of the study and the population being investigated. The study is focused on open-heart surgery and myocardial reperfusion, which are procedures related to the cardiovascular system. The complement system, which is the focus of the study, is a part of the immune system that plays a crucial role in the inflammatory response to tissue injury, including myocardial infarction. Therefore, it is relevant to study the behavior of complement activation products in the context of cardiovascular diseases. Neoplasms, on the other hand, are abnormal growths of tissue that can occur in many different organs and systems of the body. While the complement system can also play a role in the immune response to neoplasms, it is not the primary focus of this study. In summary, the study you provided is focused on cardiovascular diseases because of the nature of the procedures being investigated and the role of the complement system in the inflammatory response to tissue injury in the cardiovascular system.

1. The explanation is **useful** in distinguishing between *general pathological conditions* and *neoplasms*.

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

2. The explanation is **readable** to me (I can read it easily with comprehension).

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

3. The explanation provides **relevant** information (based on the text).

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

A.3 IMDB Movie Review Dataset - Concrete Example

Scenario

Movie staff is testing a new AI system that is supposed to explain what sentiment a movie review has, positive or negative. It does this by a so-called contrastive explanation. Those explanations aim to explain why a specific sentiment was chosen over another. The system outputs five different explanations, which you must evaluate.

Text The AI system has been provided with the following text:

It is a pity that you cannot vote zero stars on IMDb, because I would not have hesitated! In fact I would go so far as to say that this film was in the negative stars.
I, like many others, bought this film thinking that because it has Michael Madsen in it, it could be good... No chance! This film was shocking! Imagine a movie length 'The Bold and the Beautiful', well, Primal Instinct did not even come close to that good, and I had previously thought that there would be nothing worse than a movie length 'The Bold and the Beautiful'.
Michael Madsen, how could you do this to us? The worst part is, I didn't fast forward a bit, I was hoping that at the end they would reveal that it was all some sort of sick joke, that they thought it would be funny to make us watch such a horribly bad film.
Where do I start...? Directing - Zero Stars, Screenplay - Zero Stars, Acting - Zero Stars, Cinematography - Zero Stars, Digital Effects - Zero Stars, Production Design - Zero Stars, Make-up - Zero Stars, Casting - Zero Stars, Editing - Zero Stars, Trailer - Half a Star, Graphic Design - Half a Star, DVD Menu - Half a Star.
However I think that it is very important to have seen bad films just so that you know what a really bad film is, so for that reason I am happy that I saw this film, just so that I have a bad film to put at the bottom of my list.

Based on the text above, you are given 5 explanations provided by different AI systems. Answer all 3 questions provided with each explanation.

Explanation 1

The text is negative because the reviewer expresses strong dislike for the film, using language such as "in the negative stars," "shocking," "horribly bad," "worst part," and "Zero Stars" to describe it. They also mention that they didn't fast forward the film, even though they had the opportunity to, which suggests that they were not enjoying it. Additionally, the reviewer goes out of their way to list every aspect of the film that they disliked, from the directing to the casting, and even gives half a star to the trailer and DVD menu. Overall, the tone of the text is one of disappointment and frustration with the film.

1. The explanation is **useful** in distinguishing between *negative* and *positive*.

1	2	3	4	5
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

2. The explanation is **readable** to me (I can read it easily with comprehension).

1	2	3	4	5
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

3. The explanation provides **relevant** information (based on the text).

1	2	3	4	5
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

Explanation 2

This text is negative because it expresses strong disapproval and disappointment with the film "Primal Instinct". The reviewer uses strong language to describe their dislike, calling the film "shocking", "horribly bad", and stating that it's "in the negative stars". They also criticize every aspect of the film, from the directing and screenplay to the acting, cinematography, and even the DVD menu.
The reviewer's tone is sarcastic and critical, with phrases like "Michael Madsen, how could you do this to us?" and "I didn't fast forward a bit, I was hoping that at the end they would reveal that it was all some sort of sick joke". This suggests that the reviewer is not just disappointed, but also feels betrayed and frustrated by the film's poor quality.
The only slightly positive comments are the reviewer's acknowledgement that seeing bad films can be important for understanding what a really bad film is, and their willingness to give the film's trailer, graphic design, and DVD menu half a star. However, these positive comments are minor and do not outweigh the overall negative tone of the review.

1. The explanation is **useful** in distinguishing between *negative* and *positive*.

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

2. The explanation is **readable** to me (I can read it easily with comprehension).

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

3. The explanation provides **relevant** information (based on the text).

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

Explanation 3

This text is negative for several reasons:

1. Strong negative language: The text uses strong negative words and phrases, such as "pity", "shocking", "horribly bad", "zero stars", and "worst part".

2. Disappointment and frustration: The writer expresses disappointment and frustration with the film, feeling that it did not live up to their expectations, especially given the presence of Michael Madsen.

3. Comparison to a low standard: The writer compares the film unfavorably to a soap opera ("The Bold and the Beautiful"), implying that the film is even worse than something they already consider to be of low quality.

4. Sarcastic tone: The text has a sarcastic tone, with the writer saying they would have liked to give the film negative stars and implying that the filmmakers must have been joking to produce something so bad.

5. Extensive criticism: The writer provides a detailed list of aspects of the film that they consider to be of zero quality, leaving no aspect of the film unscathed.

The only slightly positive comment in the text is the writer's statement that they are "happy" to have seen the film, but this is only because it provides a baseline for comparison to other bad films. Overall, the text is overwhelmingly negative.

1. The explanation is **useful** in distinguishing between *negative* and *positive*.

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

2. The explanation is **readable** to me (I can read it easily with comprehension).

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

3. The explanation provides **relevant** information (based on the text).

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

Explanation 4

The text is negative because the author expresses strong dislike and disappointment towards the film "Primal Instinct." They use harsh language such as "negative stars," "shocking," "horribly bad," "zero stars," and "the worst part" to describe their experience with the film. They also give the film zero stars in every category, including directing, screenplay, acting, cinematography, digital effects, production design, make-up, casting, editing, trailer, graphic design, and DVD menu. The author also states that they are happy to have seen the film because it allows them to know what a really bad film is and to put it at the bottom of their list.

1. The explanation is **useful** in distinguishing between *negative* and *positive*.

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

2. The explanation is **readable** to me (I can read it easily with comprehension).

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

3. The explanation provides **relevant** information (based on the text).

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

Explanation 5

The text is negative because it expresses a strongly negative opinion about a film, calling it "shocking," "horribly bad," and "the worst part" of Michael Madsen's career. The author also gives zero stars to various aspects of the film such as directing, screenplay, acting, cinematography, digital effects, production design, makeup, casting, editing, and even the DVD menu. The author also says that they regret buying the film and that they hope it was all a sick joke. The only positive thing the author says about the film is that it serves as a good example of a really bad film, which they can use as a reference point for comparing other films. Overall, the tone and content of the text are overwhelmingly negative.

1. The explanation is **useful** in distinguishing between *negative* and *positive*.

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

2. The explanation is **readable** to me (I can read it easily with comprehension).

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

3. The explanation provides **relevant** information (based on the text).

1	**2**	**3**	**4**	**5**
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

B Detailed Results

This Appendix section includes results from the experiment without explicitly visualizing them using plots or graphs. This is so-called "raw" aggregated data in the form of means and standard deviations. The data also contains minimum and maximum values.

It is important to note that the data are based on correctly classified samples, meaning, the number of samples used for each model/dataset may differ (based on the model accuracy shown in Figure 13). Table 2 shows the number of samples used for each model and each dataset. Mistral is one of the models that has the lowest number of accurately predicted samples, with only 22 out of 100 for the medical dataset.

Table 2: Number of samples used for running evaluation for each model on each of the datasets. The number is based on 100 samples used to check model accuracy in predicting the correct label. Only the samples correctly classified were considered for the evaluation.

Model	IMDB	Medical
Llama2-7B	96	95
Llama3-8B	94	84
Llama3-70B	96	89
Mistral-7B	74	22
Mixtral-7x8B	94	79

B.1 Usefulness

Table 3: BERTScore Semantic Similarity F1 metric mean and standard deviation on the IMDB dataset.

Model	**Feature**	**Mean**	**Std**	**Min**	**Max**
Llama2-7B	Fact - CE	0.306	0.030	0.226	0.359
	NCE (Fact) - Fact	0.299	0.027	0.223	0.352
	Fact - Text	0.242	0.028	0.160	0.310
	Foil - CE	0.278	0.032	0.196	0.358
	NCE (Foil) - Foil	0.278	0.027	0.220	0.352
	Foil - Text	0.237	0.027	0.163	0.297
Llama-3-8B	Fact - CE	0.310	0.023	0.264	0.355
	NCE (Fact) - Fact	0.327	0.024	0.277	0.385
	Fact - Text	0.242	0.028	0.160	0.310
	Foil - CE	0.281	0.025	0.234	0.341
	NCE (Foil) - Foil	0.305	0.026	0.255	0.373
	Foil - Text	0.238	0.027	0.163	0.297
Llama3-70B	Fact - CE	0.289	0.019	0.253	0.335
	NCE (Fact) - Fact	0.305	0.024	0.240	0.367
	Fact - Text	0.242	0.027	0.160	0.310
	Foil - CE	0.263	0.019	0.218	0.310
	NCE (Foil) - Foil	0.298	0.024	0.243	0.350
	Foil - Text	0.237	0.027	0.163	0.297
Mistral-7B	Fact - CE	0.342	0.025	0.277	0.414
	NCE (Fact) - Fact	0.350	0.028	0.282	0.423
	Fact - Text	0.242	0.027	0.160	0.294
	Foil - CE	0.315	0.025	0.255	0.359
	NCE (Foil) - Foil	0.338	0.021	0.285	0.388
	Foil - Text	0.236	0.025	0.167	0.283
Mixtral-8x7B	Fact - CE	0.323	0.022	0.276	0.372
	NCE (Fact) - Fact	0.327	0.020	0.275	0.382
	Fact - Text	0.243	0.027	0.160	0.310
	Foil - CE	0.293	0.030	0.237	0.364
	NCE (Foil) - Foil	0.323	0.023	0.267	0.373
	Foil - Text	0.237	0.027	0.163	0.297

Table 4: BERTScore Semantic Similarity F1 metric mean and standard deviation on the Medical dataset.

Model	**Feature**	**Mean**	**Std**	**Min**	**Max**
Llama2-7B	Fact - CE	0.382	0.036	0.305	0.482
	NCE (Fact) - Fact	0.421	0.045	0.325	0.534
	Fact - Text	0.361	0.049	0.260	0.470
	Foil - CE	0.381	0.037	0.293	0.498
	NCE (Foil) - Foil	0.432	0.036	0.344	0.514
	Foil - Text	0.346	0.052	0.256	0.482
Llama3-8B	Fact - CE	0.411	0.040	0.333	0.510
	NCE (Fact) - Fact	0.454	0.047	0.372	0.565
	Fact - Text	0.355	0.047	0.260	0.470
	Foil - CE	0.400	0.035	0.328	0.473
	NCE (Foil) - Foil	0.454	0.032	0.408	0.531
	Foil - Text	0.341	0.049	0.256	0.452
Llama3-70B	Fact - CE	0.406	0.036	0.327	0.489
	NCE (Fact) - Fact	0.444	0.045	0.369	0.572
	Fact - Text	0.359	0.049	0.260	0.470
	Foil - CE	0.404	0.034	0.341	0.503
	NCE (Foil) - Foil	0.457	0.036	0.389	0.577
	Foil - Text	0.346	0.054	0.256	0.483
Mistral-7B	Fact - CE	0.460	0.042	0.392	0.551
	NCE (Fact) - Fact	0.478	0.049	0.379	0.607
	Fact - Text	0.366	0.048	0.281	0.458
	Foil - CE	0.470	0.033	0.421	0.539
	NCE (Foil) - Foil	0.474	0.054	0.349	0.568
	Foil - Text	0.347	0.056	0.256	0.452
Mixtral-8x7B	Fact - CE	0.417	0.038	0.331	0.499
	NCE (Fact) - Fact	0.432	0.035	0.361	0.515
	Fact - Text	0.356	0.047	0.260	0.470
	Foil - CE	0.404	0.043	0.321	0.508
	NCE (Foil) - Foil	0.441	0.034	0.371	0.527
	Foil - Text	0.340	0.051	0.256	0.483

B.2 Generality

B.2.1 Lexical Diversity

Table 5: Lexical Diversity - the MATTR (window size = 50) score mean, standard deviation, min and max for IMDB dataset.

Model	Feature	Mean	Std	Min	Max
Llama2-7B	Text (MATTR)	0.816	0.035	0.725	0.931
	Modified Text (MATTR)	0.812	0.035	0.721	0.934
	Text (Word Count)	217.000	162.281	48.000	800.000
	Modified Text (Word Count)	220.042	164.466	49.000	815.000
	CE (MATTR)	0.746	0.043	0.657	0.856
	MCE (MATTR)	0.744	0.048	0.607	0.865
	CE (Word Count)	143.562	57.331	53.000	321.000
	MCE (Word Count)	137.771	51.709	65.000	269.000
Llama3-8B	Text (MATTR)	0.815	0.034	0.725	0.931
	Modified Text (MATTR)	0.811	0.035	0.721	0.934
	Text (Word Count)	211.415	156.335	48.000	800.000
	Modified Text (Word Count)	214.351	158.491	49.000	815.000
	CE (MATTR)	0.756	0.042	0.638	0.840
	MCE (MATTR)	0.757	0.039	0.644	0.866
	CE (Word Count)	143.521	33.505	74.000	278.000
	MCE (Word Count)	150.681	30.145	75.000	212.000
Llama3-70B	Text (MATTR)	0.816	0.035	0.725	0.931
	Modified Text (MATTR)	0.812	0.035	0.721	0.934
	Text (Word Count)	211.552	155.234	48.000	800.000
	Modified Text (Word Count)	214.510	157.372	49.000	815.000
	CE (MATTR)	0.763	0.029	0.688	0.832
	MCE (MATTR)	0.761	0.034	0.671	0.832
	CE (Word Count)	175.281	26.342	115.000	235.000
	MCE (Word Count)	183.292	28.425	123.000	268.000
Mistral-7B	Text (MATTR)	0.815	0.037	0.725	0.931
	Modified Text (MATTR)	0.811	0.037	0.721	0.934
	Text (Word Count)	212.149	153.808	55.000	800.000
	Modified Text (Word Count)	215.189	156.059	55.000	815.000
	CE (MATTR)	0.763	0.058	0.638	0.894
	MCE (MATTR)	0.761	0.058	0.647	0.875
	CE (Word Count)	74.324	23.982	41.000	197.000
	MCE (Word Count)	72.703	30.466	40.000	240.000
Mixtral-8x7B	Text (MATTR)	0.816	0.035	0.725	0.931
	Modified Text (MATTR)	0.812	0.036	0.721	0.934
	Text (Word Count)	210.404	155.865	48.000	800.000
	Modified Text (Word Count)	213.404	158.040	49.000	815.000
	CE (MATTR)	0.762	0.040	0.661	0.858
	MCE (MATTR)	0.761	0.051	0.634	0.919
	CE (Word Count)	112.840	37.952	54.000	279.000
	MCE (Word Count)	114.872	42.641	53.000	314.000

Table 6: Lexical Diversity - the MATTR (window size = 50) score mean, standard deviation, min and max for Medical dataset.

Model	**Feature**	**Mean**	**Std**	**Min**	**Max**
Llama2-7B	Text (MATTR)	0.766	0.053	0.645	0.891
	Modified Text (MATTR)	0.758	0.054	0.625	0.893
	Text (Word Count)	170.558	71.946	43.000	366.000
	Modified Text (Word Count)	174.063	74.006	43.000	371.000
	CE (MATTR)	0.743	0.037	0.654	0.821
	MCE (MATTR)	0.744	0.038	0.633	0.835
	CE (Word Count)	160.274	45.763	68.000	314.000
	MCE (Word Count)	165.716	46.021	71.000	298.000
Llama3-8B	Text (MATTR)	0.763	0.051	0.645	0.888
	Modified Text (MATTR)	0.756	0.053	0.625	0.893
	Text (Word Count)	181.179	74.552	44.000	398.000
	Modified Text (Word Count)	184.976	76.687	45.000	407.000
	CE (MATTR)	0.753	0.040	0.651	0.867
	MCE (MATTR)	0.748	0.044	0.625	0.847
	CE (Word Count)	135.631	32.738	48.000	224.000
	MCE (Word Count)	143.107	28.003	86.000	220.000
Llama3-70B	Text (MATTR)	0.766	0.052	0.645	0.888
	Modified Text (MATTR)	0.759	0.055	0.625	0.893
	Text (Word Count)	172.910	76.808	43.000	398.000
	Modified Text (Word Count)	176.461	78.913	43.000	407.000
	CE (MATTR)	0.752	0.049	0.610	0.853
	MCE (MATTR)	0.758	0.050	0.604	0.857
	CE (Word Count)	116.640	28.493	55.000	189.000
	MCE (Word Count)	118.831	25.870	64.000	179.000
Mistral-7B	Text (MATTR)	0.773	0.054	0.678	0.891
	Modified Text (MATTR)	0.766	0.051	0.658	0.871
	Text (Word Count)	193.364	58.914	68.000	290.000
	Modified Text (Word Count)	197.182	60.071	69.000	292.000
	CE (MATTR)	0.748	0.057	0.648	0.866
	MCE (MATTR)	0.764	0.063	0.613	0.857
	CE (Word Count)	72.364	20.779	45.000	133.000
	MCE (Word Count)	88.773	34.451	28.000	170.000
Mixtral-8x7B	Text (MATTR)	0.767	0.052	0.645	0.888
	Modified Text (MATTR)	0.760	0.054	0.625	0.893
	Text (Word Count)	180.646	74.147	53.000	398.000
	Modified Text (Word Count)	184.051	76.036	53.000	407.000
	CE (MATTR)	0.775	0.046	0.675	0.966
	MCE (MATTR)	0.767	0.041	0.672	0.875
	CE (Word Count)	132.759	47.217	60.000	265.000
	MCE (Word Count)	127.785	56.240	8.000	338.000

B.2.2 Lexical Density

Table 7: Lexical Density (the ratio of content words over total words) mean and standard deviation for IMDB dataset.

Model	Feature	Mean	Std	Min	Max
Llama2-7B	Text Density	0.404	0.045	0.229	0.508
	Mofidied Text Density	0.391	0.046	0.226	0.500
	CE Density	0.415	0.027	0.351	0.479
	MCE Density	0.413	0.025	0.354	0.490
Llama3-8B	Text Density	0.406	0.041	0.238	0.508
	Mofidied Text Density	0.393	0.043	0.238	0.500
	CE Density	0.411	0.027	0.347	0.489
	MCE Density	0.410	0.031	0.338	0.503
Llama-3-70B	Text Density	0.404	0.045	0.229	0.508
	Mofidied Text Density	0.391	0.046	0.226	0.500
	CE Density	0.413	0.020	0.336	0.465
	MCE Density	0.410	0.022	0.348	0.475
Mistral-7B	Text Density	0.407	0.038	0.328	0.508
	Mofidied Text Density	0.393	0.040	0.309	0.500
	CE Density	0.417	0.040	0.303	0.487
	MCE Density	0.412	0.041	0.314	0.516
Mixtral-8x7B	Text Density	0.404	0.045	0.229	0.508
	Mofidied Text Density	0.391	0.046	0.226	0.500
	CE Density	0.434	0.030	0.361	0.493
	MCE Density	0.428	0.031	0.353	0.511

Table 8: Lexical Density (the ratio of content words over total words) mean and standard deviation for Medical dataset.

Model	Feature	Mean	Std	Min	Max
Llama2-7B	Text Density	0.505	0.062	0.355	0.638
	Modified Text Density	0.500	0.064	0.355	0.630
	CE Density	0.480	0.025	0.422	0.533
	MCE Density	0.477	0.027	0.413	0.532
Llama3-8B	Text Density	0.503	0.059	0.355	0.603
	Modified Text Density	0.497	0.060	0.355	0.603
	CE Density	0.476	0.027	0.392	0.548
	MCE Density	0.474	0.030	0.376	0.546
Llama3-70B	Text Density	0.505	0.059	0.355	0.625
	Modified Text Density	0.500	0.060	0.355	0.625
	CE Density	0.453	0.033	0.350	0.553
	MCE Density	0.464	0.031	0.370	0.545
Mistral-7B	Text Density	0.500	0.063	0.369	0.638
	Modified Text Density	0.496	0.061	0.366	0.630
	CE Density	0.489	0.029	0.430	0.533
	MCE Density	0.503	0.039	0.431	0.585
Mixtral-8x7B	Text Density	0.504	0.059	0.369	0.605
	Modified Text Density	0.499	0.061	0.366	0.603
	CE Density	0.493	0.032	0.427	0.573
	MCE Density	0.485	0.035	0.300	0.572

B.2.3 Semantic Similarity

Table 9: Sentence-BERT Semantic Similarity scores between Text and Modified Text, as well as Contrastive Explanation and Modified Contrastive Explanation for IMDB dataset.

Model	Feature	Mean	Std	Min	Max
Llama2-7B	CE - MCE	0.772	0.098	0.449	0.969
	Text - Modified Text	0.766	0.096	0.419	0.935
Llama3-8B	CE - MCE	0.815	0.090	0.623	0.981
	Text - Modified Text	0.767	0.094	0.419	0.935
Llama3-70B	CE - MCE	0.836	0.073	0.645	0.983
	Text - Modified Text	0.768	0.094	0.419	0.935
Mistral-7B	CE - MCE	0.782	0.094	0.529	0.982
	Text - Modified Text	0.765	0.099	0.419	0.935
Mixtral-8x7B	CE - MCE	0.820	0.087	0.593	0.989
	Text - Modified Text	0.768	0.094	0.419	0.935

Table 10: Sentence-BERT Semantic Similarity scores between Text and Modified Text, as well as Contrastive Explanation and Modified Contrastive Explanation for Medical dataset.

Model	Feature	Mean	Std	Min	Max
Llama2-7B	CE - MCE	0.855	0.086	0.615	0.982
	Text - Modified Text	0.900	0.063	0.712	0.990
Llama3-8B	CE - MCE	0.872	0.086	0.541	0.985
	Text - Modified Text	0.901	0.062	0.712	0.990
Llama3-70B	CE - MCE	0.878	0.081	0.584	0.981
	Text - Modified Text	0.898	0.066	0.712	0.990
Mistral-7B	CE - MCE	0.832	0.107	0.607	0.980
	Text - Modified Text	0.893	0.049	0.746	0.957
Mixtral-8x7B	CE - MCE	0.842	0.111	0.391	0.970
	Text - Modified Text	0.896	0.068	0.712	0.990

B.3 Readability

Table 11: Readability measures for IMDB dataset.

Model	Feature	Mean	Std	Min	Max
Llama2-7B	Syllables/word (Text)	1.295	0.081	1.100	1.516
	Words/Sentence (Text)	21.411	7.294	5.000	51.500
	Words (Text)	253.125	191.889	59.000	954.000
	Syllables/Word (CE)	1.435	0.062	1.310	1.602
	Words/Sentence (CE)	26.737	6.131	13.300	53.333
	Words (CE)	175.427	67.591	70.000	373.000
	Syllables/Word (NCE Fact)	1.436	0.073	1.305	1.701
	Words/Sentence (NCE Fact)	24.961	4.583	13.444	36.600
	Words (NCE Fact)	192.719	62.787	88.000	362.000
	Syllables/Word (NCE Foil)	1.421	0.068	1.237	1.611
	Words/Sentence (NCE Foil)	25.810	5.742	15.615	53.000
	Words (NCE Foil)	199.000	56.064	90.000	337.000
Llama3-8B	Syllables/word (Text)	1.298	0.081	1.100	1.516
	Words/Sentence (Text)	21.429	7.386	5.000	51.500
	Words (Text)	246.372	184.188	59.000	954.000
	Syllables/Word (CE)	1.451	0.062	1.301	1.599
	Words/Sentence (CE)	25.919	7.427	12.333	63.667
	Words (CE)	176.872	41.155	91.000	363.000
	Syllables/Word (NCE Fact)	1.451	0.071	1.312	1.692
	Words/Sentence (NCE Fact)	24.678	6.519	12.083	57.000
	Words (NCE Fact)	142.053	40.247	77.000	271.000
	Syllables/Word (NCE Foil)	1.449	0.070	1.274	1.648
	Words/Sentence (NCE Foil)	20.075	7.847	8.111	46.667
	Words (NCE Foil)	143.968	35.405	63.000	246.000
Llama3-70B	Syllables/word (Text)	1.297	0.081	1.100	1.516
	Words/Sentence (Text)	21.460	7.312	5.000	51.500
	Words (Text)	246.448	182.677	59.000	954.000
	Syllables/Word (CE)	1.455	0.053	1.323	1.579
	Words/Sentence (CE)	22.587	4.497	15.588	38.500
	Words (CE)	225.292	37.073	148.000	303.000
	Syllables/Word (NCE Fact)	1.450	0.063	1.288	1.613
	Words/Sentence (NCE Fact)	25.159	6.734	11.200	55.000
	Words (NCE Fact)	192.458	41.489	69.000	300.000
	Syllables/Word (NCE Foil)	1.437	0.061	1.313	1.587
	Words/Sentence (NCE Foil)	22.984	6.391	13.091	40.500
	Words (NCE Foil)	159.958	36.004	82.000	243.000
Mistral-7B	Syllables/word (Text)	1.299	0.083	1.123	1.516
	Words/Sentence (Text)	20.829	6.471	10.842	37.143
	Words (Text)	247.014	183.377	60.000	954.000
	Syllables/Word (CE)	1.432	0.088	1.264	1.620
	Words/Sentence (CE)	29.709	12.624	14.667	74.000
	Words (CE)	92.892	29.074	44.000	235.000
	Syllables/Word (NCE Fact)	1.463	0.086	1.235	1.645
	Words/Sentence (NCE Fact)	28.404	16.188	12.091	132.000
	Words (NCE Fact)	87.649	26.718	39.000	194.000
	Syllables/Word (NCE Foil)	1.455	0.104	1.220	1.696
	Words/Sentence (NCE Foil)	32.681	24.201	9.083	153.000
	Words (NCE Foil)	86.297	31.933	35.000	231.000
Mixtral-8x7B	Syllables/word (Text)	1.295	0.081	1.100	1.516
	Words/Sentence (Text)	21.263	7.234	5.000	51.500
	Words (Text)	245.255	183.637	59.000	954.000
	Syllables/Word (CE)	1.483	0.083	1.289	1.662

Continued on next page...

Model	Feature	Mean	Std	Min	Max
	Words/Sentence (CE)	25.557	5.383	12.917	39.000
	Words (CE)	133.266	44.616	63.000	324.000
	Syllables/Word (NCE Fact)	1.487	0.094	1.305	1.740
	Words/Sentence (NCE Fact)	24.913	5.642	13.417	43.000
	Words (NCE Fact)	129.319	37.507	71.000	250.000
	Syllables/Word (NCE Foil)	1.451	0.077	1.297	1.663
	Words/Sentence (NCE Foil)	25.561	5.044	14.273	46.000
	Words (NCE Foil)	102.191	19.034	50.000	157.000

Table 12: Readability measures for Medical dataset.

Model	Feature	Mean	Std	Min	Max
Mistral-7B	Syllables/word (Text)	1.684	0.156	1.352	1.977
	Words/Sentence (Text)	26.423	7.086	15.875	43.000
	Words (Text)	245.227	85.772	82.000	365.000
	Syllables/Word (CE)	1.609	0.115	1.477	1.917
	Words/Sentence (CE)	26.964	6.172	17.000	47.500
	Words (CE)	82.773	24.341	49.000	154.000
	Syllables/Word (NCE Fact)	1.677	0.147	1.444	2.000
	Words/Sentence (NCE Fact)	24.653	3.586	17.000	31.000
	Words (NCE Fact)	102.455	33.849	30.000	183.000
	Syllables/Word (NCE Foil)	1.587	0.108	1.417	1.774
	Words/Sentence (NCE Foil)	24.455	5.883	16.000	37.000
	Words (NCE Foil)	72.318	47.230	32.000	244.000
Llama3-70B	Syllables/word (Text)	1.700	0.150	1.308	1.978
	Words/Sentence (Text)	24.414	6.844	12.714	43.000
	Words (Text)	215.079	104.560	48.000	516.000
	Syllables/Word (CE)	1.604	0.094	1.392	1.867
	Words/Sentence (CE)	32.247	6.344	18.429	47.500
	Words (CE)	138.000	33.470	64.000	222.000
	Syllables/Word (NCE Fact)	1.671	0.143	1.372	2.000
	Words/Sentence (NCE Fact)	26.430	5.616	14.000	45.500
	Words (NCE Fact)	120.787	52.575	41.000	277.000
	Syllables/Word (NCE Foil)	1.570	0.100	1.290	1.815
	Words/Sentence (NCE Foil)	31.024	7.887	14.625	59.000
	Words (NCE Foil)	82.067	24.793	37.000	168.000
Llama2-7B	Syllables/word (Text)	1.700	0.152	1.308	1.978
	Words/Sentence (Text)	24.349	6.768	12.714	43.000
	Words (Text)	211.716	98.787	48.000	481.000
	Syllables/Word (CE)	1.609	0.088	1.421	1.891
	Words/Sentence (CE)	29.840	6.563	14.875	44.500
	Words (CE)	187.263	53.504	79.000	369.000
	Syllables/Word (NCE Fact)	1.645	0.149	1.333	2.010
	Words/Sentence (NCE Fact)	29.542	5.582	16.333	44.333
	Words (NCE Fact)	163.495	70.988	49.000	344.000
	Syllables/Word (NCE Foil)	1.537	0.119	1.221	1.830
	Words/Sentence (NCE Foil)	23.337	5.232	14.000	40.000
	Words (NCE Foil)	93.474	37.007	28.000	200.000
Llama3-8B	Syllables/word (Text)	1.693	0.148	1.308	1.977
	Words/Sentence (Text)	25.101	6.644	14.167	43.000
	Words (Text)	225.655	102.469	50.000	516.000
	Syllables/Word (CE)	1.609	0.104	1.407	1.839
	Words/Sentence (CE)	30.249	5.399	19.333	51.000
	Words (CE)	156.274	36.851	58.000	250.000
	Syllables/Word (NCE Fact)	1.646	0.150	1.347	1.894

Continued on next page...

Model	Feature	Mean	Std	Min	Max
	Words/Sentence (NCE Fact)	29.027	6.896	16.000	47.000
	Words (NCE Fact)	106.214	51.028	34.000	287.000
	Syllables/Word (NCE Foil)	1.595	0.105	1.270	1.856
	Words/Sentence (NCE Foil)	30.321	5.311	19.800	52.000
	Words (NCE Foil)	71.048	13.805	31.000	100.000
Mixtral-8x7B	Syllables/word (Text)	1.702	0.143	1.336	1.978
	Words/Sentence (Text)	24.444	6.230	12.714	43.000
	Words (Text)	224.139	102.253	56.000	516.000
	Syllables/Word (CE)	1.633	0.101	1.398	1.993
	Words/Sentence (CE)	28.158	4.384	17.500	43.500
	Words (CE)	151.949	52.954	70.000	291.000
	Syllables/Word (NCE Fact)	1.644	0.122	1.365	1.959
	Words/Sentence (NCE Fact)	27.969	5.511	18.333	46.750
	Words (NCE Fact)	145.392	52.554	55.000	304.000
	Syllables/Word (NCE Foil)	1.579	0.085	1.365	1.764
	Words/Sentence (NCE Foil)	23.404	5.068	13.333	41.000
	Words (NCE Foil)	82.937	27.352	34.000	153.000

C Questionnaires

C.1 Movie Reviews

Model	Feature	Likert Scale Frequencies					Average
		1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)	
Llama2-7B	Usefulness	1	2	2	11	4	3,75
	Readability	1	0	2	7	10	4,25
	Relevance	0	1	3	9	7	4,1
Llama3-8B	Usefulness	2	1	0	6	11	4,15
	Readability	1	0	1	3	15	4,55
	Relevance	1	0	1	4	14	4,5
Llama3-70B	Usefulness	1	1	0	7	11	4,3
	Readability	0	1	1	3	15	4,6
	Relevance	1	0	2	4	13	4,4
Mistral-7B	Usefulness	0	3	3	7	7	3,9
	Readability	1	0	4	7	8	4,05
	Relevance	1	0	3	10	6	4
Mixtral-8x7B	Usefulness	1	2	2	12	3	3,7
	Readability	1	0	1	10	8	4,2
	Relevance	0	3	5	10	2	3,55

Table 13: Movie reviews questionnaire results based on 20 responses. Likert Scale Frequencies are present for all 5 scales and average score is included.

C.2 Medical Abstracts

Model	Feature	Likert Scale Frequencies					Average
		1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)	
Llama2-7B	Usefulness	0	1	0	6	3	4,1
	Readability	0	0	2	3	5	4,3
	Relevance	1	5	1	3	0	2,6
Llama3-8B	Usefulness	0	1	0	4	5	4,3
	Readability	0	0	0	3	7	4,7
	Relevance	0	0	0	7	3	4,3
Llama3-70B	Usefulness	0	2	1	4	3	3,8
	Readability	1	4	1	2	2	3
	Relevance	1	3	3	0	3	3,1
Mistral-7B	Usefulness	8	2	0	0	0	1,2
	Readability	0	3	0	2	5	3,9
	Relevance	1	3	1	3	2	3,2
Mixtral-8x7B	Usefulness	0	3	1	2	4	3,7
	Readability	2	1	0	4	3	3,5
	Relevance	1	0	0	5	4	4,1

Table 14: Medical abstracts questionnaire results based on 10 responses. Likert Scale Frequencies are present for all 5 scales and average score is included.

D Prompts Used in the CoSEM Pipeline

D.1 The IMDB Dataset - Example

Text Well...tremors I, the original started off in 1990 and i found the movie quite enjoyable to watch. however, they proceeded to make tremors II and III. Trust me, those movies started going downhill right after they finished the first one, i mean, ass blasters??? Now, only God himself is capable of answering the question "why in Gods name would they create another one of these dumpster dives of a movie?" Tremors IV cannot be considered a bad movie, in fact it cannot be even considered an epitome of a bad movie, for it lives up to more than that. As i attempted to sit though it, i noticed that my eyes started to bleed, and i hoped profusely that the little girl from the ring would crawl through the TV and kill me. did they really think that dressing the people who had stared in the other movies up as though they we're from the wild west would make the movie (with the exact same occurrences) any better? honestly, i would never suggest buying this movie, i mean, there are cheaper ways to find things that burn well.

Fact negative

Foil positive

D.1.1 Model Accuracy

What label does the following text belong to: negative or positive? If you can't tell, then "unsure". Output only one word.

Well...tremors I, the original started off in 1990 and i found the movie quite enjoyable to watch. however, they proceeded to make tremors II and III. Trust me, those movies started going downhill right after they finished the first one, i mean, ass blasters??? Now, only God himself is capable of answering the question "why in Gods name would they create another one of these dumpster dives of a movie?" Tremors IV cannot be considered a bad movie, in fact it cannot be even considered an epitome of a bad movie, for it lives up to more than that. As i attempted to sit though it, i noticed that my eyes started to bleed, and i hoped profusely that the little girl from the ring would crawl through the TV and kill me. did they really think that dressing the people who had stared in the other movies up as though they we're from the wild west would make the movie (with the exact same occurrences) any better? honestly, i would never suggest buying this movie, i mean, there are cheaper ways to find things that burn well.

Negative

Figure 28: Example of prompting `meta-llama/Meta-Llama-3-8B-Instruct` model on the movie reviews dataset to predict the sentiment for a text sample. The model responds with either positive, negative or unsure. In this case, the model correctly predicted that the sentiment was negative.

D.1.2 Model Self-Consistency

For the following text, why is it negative and not positive?

Well...tremors I, the original started off in 1990 and i found the movie quite enjoyable to watch. however, they proceeded to make tremors II and III. Trust me, those movies started going downhill right after they finished the first one, i mean, ass blasters??? Now, only God himself is capable of answering the question "why in Gods name would they create another one of these dumpster dives of a movie?" Tremors IV cannot be considered a bad movie, in fact it cannot be even considered an epitome of a bad movie, for it lives up to more than that. As i attempted to sit though it, i noticed that my eyes started to bleed, and i hoped profusely that the little girl from the ring would crawl through the TV and kill me. did they really think that dressing the people who had stared in the other movies up as though they we're from the wild west would make the movie (with the exact same occurrences) any better? honestly, i would never suggest buying this movie, i mean, there are cheaper ways to find things that burn well.

The review is negative because:
1. The reviewer uses strong language to express their dislike, such as "dumpster dive", "bleed", and "kill me". 2. They criticize the sequels, calling them "going downhill" and stating that they can't understand why anyone would want to make another one. 3. They specifically mention what they didn't like about each sequel, including the addition of "ass blasters" and the Western-themed costumes. 4. They conclude by saying they wouldn't recommend buying the movie, implying that it's not worth spending money on.
Overall, the tone is sarcastic and critical, with the reviewer expressing extreme disappointment and frustration with the series.

Figure 29: Prompting `meta-llama/Meta-Llama-3-8B-Instruct` model on the movie reviews dataset to generate a contrastive self-explanation ("Why A and not B?"). The model is prompted to generate such a self-explanation only when it correctly predicts the label (Figure 28).

In the text below, redact the words necessary to answer question why negative and not positive. Replace important words with [REDACTED]. Do not explain the answer. Give only redacted text.

Well...tremors I, the original started off in 1990 and i found the movie quite enjoyable to watch. however, they proceeded to make tremors II and III. Trust me, those movies started going downhill right after they finished the first one, i mean, ass blasters??? Now, only God himself is capable of answering the question "why in Gods name would they create another one of these dumpster dives of a movie?" Tremors IV cannot be considered a bad movie, in fact it cannot be even considered an epitome of a bad movie, for it lives up to more than that. As i attempted to sit though it, i noticed that my eyes started to bleed, and i hoped profusely that the little girl from the ring would crawl through the TV and kill me. did they really think that dressing the people who had stared in the other movies up as though they we're from the wild west would make the movie (with the exact same occurrences) any better? honestly, i would never suggest buying this movie, i mean, there are cheaper ways to find things that burn well.

Here is the redacted text:
Well...[REDACTED] I, the original started off in 1990 and i found the movie quite enjoyable to watch. However, they proceeded to make [REDACTED] II and III. Trust me, those movies started going [REDACTED] right after they finished the first one, i mean, [REDACTED]? Now, only [REDACTED] himself is capable of answering the question "[REDACTED]"? [REDACTED] IV cannot be considered a [REDACTED] movie, in fact it cannot be even considered an [REDACTED] of a [REDACTED], for it lives up to more than that. As i attempted to sit though it, i noticed that my eyes started to [REDACTED], and i hoped profusely that the little girl from the [REDACTED] would crawl through the TV and kill me. Did they really think that dressing the people who had starred in the other movies up as though they were from the [REDACTED] would make the movie (with the exact same [REDACTED]) any better? Honestly, i would never suggest buying this movie, i mean, there are cheaper ways to find things that [REDACTED].

Figure 30: Example of prompting the `meta-llama/Meta-Llama-3-8B-Instruct` model on the movie reviews dataset to output a redacted text which does not allow to explain both labels. This is an extension of the method from [39]. The model masks tokens with [REDACTED] that it considers crucial for answering the contrastive question. This redaction is applied additional $k = 5$ times (6 in total) for quantifiable results.

What is the class of the following text? The text can contain redacted words marked with [REDACTED]. Answer only negative, positive, or unsure. Do not explain the answer.

Well...[REDACTED] I, the original started off in 1990 and i found the movie quite enjoyable to watch. However, they proceeded to make [REDACTED] II and III. Trust me, those movies started going [REDACTED] right after they finished the first one, i mean, [REDACTED]? Now, only [REDACTED] himself is capable of answering the question "[REDACTED]"? [REDACTED] IV cannot be considered a [REDACTED] movie, in fact it cannot be even considered an [REDACTED] of a [REDACTED], for it lives up to more than that. As i attempted to sit though it, i noticed that my eyes started to [REDACTED], and i hoped profusely that the little girl from the [REDACTED] would crawl through the TV and kill me. Did they really think that dressing the people who had starred in the other movies up as though they were from the [REDACTED] would make the movie (with the exact same [REDACTED]) any better? Honestly, i would never suggest buying this movie, i mean, there are cheaper ways to find things that [REDACTED].

Negative

Figure 31: Example of prompting the `meta-llama/Meta-Llama-3-8B-Instruct` to predict the label on the redacted text. If the model predicts "unsure", then it is self-consistent. In any other case, it is not self-consistent as it claimed to redact words necessary for it to predict the label (answer the contrastive question). In this case, the model is not self-consistent, as it replied "negative".

D.1.3 Explanation Generation

For the following text, why is it negative?

Well...tremors I, the original started off in 1990 and i found the movie quite enjoyable to watch. however, they proceeded to make tremors II and III. Trust me, those movies started going downhill right after they finished the first one, i mean, ass blasters??? Now, only God himself is capable of answering the question "why in Gods name would they create another one of these dumpster dives of a movie?" Tremors IV cannot be considered a bad movie, in fact it cannot be even considered an epitome of a bad movie, for it lives up to more than that. As i attempted to sit though it, i noticed that my eyes started to bleed, and i hoped profusely that the little girl from the ring would crawl through the TV and kill me. did they really think that dressing the people who had stared in the other movies up as though they we're from the wild west would make the movie (with the exact same occurrences) any better? honestly, i would never suggest buying this movie, i mean, there are cheaper ways to find things that burn well.

The tone of the text is overwhelmingly negative because:
1. The author describes the sequels ("Tremors II" and "III") as "dumpster dives", implying they are worthless or useless. 2. They express frustration and disappointment with the direction the franchise took after the first film. 3. The use of hyperbole ("my eyes started to bleed") emphasizes their strong dislike for the fourth installment. 4. The comparison to the Ring's cursed videotape implies that watching Tremors IV was almost as traumatic as experiencing something supernatural. 5. The final sentence suggests that not even burning money could be worse than wasting time on this movie.
Overall, the language used is sarcastic, critical, and dismissive, indicating a strongly negative opinion about the quality of the films.

Figure 32: Example of prompting the `meta-llama/Meta-Llama-3-8B-Instruct` model on the movie reviews dataset to generate non-contrastive self-explanation NCE for the fact label.

For the following text, why is it NOT positive?

Well...tremors I, the original started off in 1990 and i found the movie quite enjoyable to watch. however, they proceeded to make tremors II and III. Trust me, those movies started going downhill right after they finished the first one, i mean, ass blasters??? Now, only God himself is capable of answering the question "why in Gods name would they create another one of these dumpster dives of a movie?" Tremors IV cannot be considered a bad movie, in fact it cannot be even considered an epitome of a bad movie, for it lives up to more than that. As i attempted to sit though it, i noticed that my eyes started to bleed, and i hoped profusely that the little girl from the ring would crawl through the TV and kill me. did they really think that dressing the people who had stared in the other movies up as though they we're from the wild west would make the movie (with the exact same occurrences) any better? honestly, i would never suggest buying this movie, i mean, there are cheaper ways to find things that burn well.

This text is not positive because:
* The writer describes the sequels to the original movie ("Tremors II" and "III") as being terrible and "dumpster dives". * They express frustration and disappointment with the direction the series took. * They use strong negative language to describe their experience watching "Tremors IV", saying their eyes began to bleed and wishing someone would come and kill them. * They explicitly state that they would not recommend buying or watching the movie. Overall, the tone of the text is overwhelmingly critical and negative towards the movie franchise.

Figure 33: Example of prompting the `meta-llama/Meta-Llama-3-8B-Instruct` model on the movie reviews dataset to generate non-contrastive self-explanation NCE for the foil label.

For the following text, why is it negative and not positive?

Well... tremor I, the creation started off in 1990 and i found the show quite enjoyable to watch. however, they proceeded to make tremor II and III. Trust me, those show started going downhill right after they finished the first digit, i mean, butt blaster??? Now, only God himself is capable of answering the questioning " why in Gods name would they create another digit of these dumpster cabaret of a show? " tremor digit can not be considered a bad show, in information it can not be even considered an epitome of a bad show, for it lives up to more than that. As i attempted to sit though it, i noticed that my opinion started to bleed, and i hoped profusely that the small woman from the sound would crawl through the broadcasting and kill me. did they really think that dressing the group who had stared in the other show up as though they we 're from the wild west would make the show (with the exact same event) any better? honestly, i would never suggest buying this show, i mean, there are inexpensive structure to find property that burn well.

The review is negative because:
1. The reviewer states that the sequels ("Tremors II" and "III") were worse than the original. 2. They use strong language to express their dislike, such as calling them a "dumpster cabaret", implying that they're worthless or trashy. 3. They question the creators' sanity by asking "why in God's name..." which implies confusion and frustration. 4. They mention feeling physically uncomfortable while watching the movie, saying their opinion "started to bleed" and hoping someone would come and rescue them. 5. They explicitly state that they wouldn't recommend buying the show, comparing it unfavorably to something else that burns well. Overall, the tone is overwhelmingly critical and dismissive, indicating a strongly negative view of the movies.

Figure 34: Prompting `meta-llama/Meta-Llama-3-8B-Instruct` model on the movie reviews dataset to generate a modified contrastive self-explanation ("Why A and not B?"). The explanation is later evaluated on generality with the original CE (Figure 29).

D.2 The Medical Dataset - Example

Text Tissue changes around loose prostheses. A canine model to investigate the effects of an antiinflammatory agent. The aseptically loosened prosthesis provided a means for investigating the in vivo and in vitro activity of the cells associated with the loosening process in seven dogs. The cells were isolated and maintained in culture for sufficient periods of time so that their biologic activity could be studied as well as the effect of different agents added to the cells in vivo or in vitro. The biologic response as determined by interleukin-1 and prostaglandin E2 activity paralleled the roentgenographic appearance of loosening and the technetium images and observations made at the time of revision surgery. The correlation between clinical, roentgenographic, histologic, and biochemical loosening indicates that the canine model is suitable for investigating the mechanisms of prosthetic failure. A canine model permits the study of possible nonsurgical therapeutic interventions with the ultimate hope of stopping or slowing the loosening process.

Fact general pathological conditions

Foil neoplasms

D.2.1 Model Accuracy

What label does the following text belong to: general pathological conditions or neoplasms? If you can't tell, then "unsure". Output only one word.

Tissue changes around loose prostheses. A canine model to investigate the effects of an antiinflammatory agent. The aseptically loosened prosthesis provided a means for investigating the in vivo and in vitro activity of the cells associated with the loosening process in seven dogs. The cells were isolated and maintained in culture for sufficient periods of time so that their biologic activity could be studied as well as the effect of different agents added to the cells in vivo or in vitro. The biologic response as determined by interleukin-1 and prostaglandin E2 activity paralleled the roentgenographic appearance of loosening and the technetium images and observations made at the time of revision surgery. The correlation between clinical, roentgenographic, histologic, and biochemical loosening indicates that the canine model is suitable for investigating the mechanisms of prosthetic failure. A canine model permits the study of possible nonsurgical therapeutic interventions with the ultimate hope of stopping or slowing the loosening process.

General pathological conditions

Figure 35: Example of prompting `meta-llama/Meta-Llama-3-8B-Instruct` model on the medical dataset to predict the sentiment for a text sample. The model responds with either positive, negative or unsure. In this case, the model correctly predicted that the sentiment was negative.

D.2.2 Model Self-Consistency

For the following text, why is it general pathological conditions and not neoplasms?

Tissue changes around loose prostheses. A canine model to investigate the effects of an antiinflammatory agent. The aseptically loosened prosthesis provided a means for investigating the in vivo and in vitro activity of the cells associated with the loosening process in seven dogs. The cells were isolated and maintained in culture for sufficient periods of time so that their biologic activity could be studied as well as the effect of different agents added to the cells in vivo or in vitro. The biologic response as determined by interleukin-1 and prostaglandin E2 activity paralleled the roentgenographic appearance of loosening and the technetium images and observations made at the time of revision surgery. The correlation between clinical, roentgenographic, histologic, and biochemical loosening indicates that the canine model is suitable for investigating the mechanisms of prosthetic failure. A canine model permits the study of possible nonsurgical therapeutic interventions with the ultimate hope of stopping or slowing the loosening process.

The reason this research focuses on ""general pathological conditions"" rather than neoplasms (cancer) is because the topic being investigated is related to tissue damage and inflammation caused by implant loosening, which is a common problem in orthopedic implants such as hip replacements.
Loose prostheses can lead to chronic inflammation, bone loss, and eventual device failure, but these issues are not directly related to cancer development. Instead, they involve complex interactions between immune cells, inflammatory mediators, and mechanical factors that contribute to tissue degradation over time.
By studying the cellular responses and biological processes involved in implant loosening, researchers aim to develop new treatments or therapies to prevent or slow down this condition, ultimately improving patient outcomes without necessarily targeting cancer-specific pathways or biomarkers. Therefore, the focus is on understanding and addressing non-neoplastic diseases or disorders affecting joint replacement devices.

Figure 36: Prompting `meta-llama/Meta-Llama-3-8B-Instruct` model on the medical dataset to generate a contrastive self-explanation ("Why A and not B?"). The model is prompted to generate such a self-explanation only when it correctly predicts the label (Figure 35).

In the text below, redact the words necessary to answer question why general pathological conditions and not neoplasms. Replace important words with [REDACTED]. Do not explain the answer. Give only redacted text.

Tissue changes around loose prostheses. A canine model to investigate the effects of an antiinflammatory agent. The aseptically loosened prosthesis provided a means for investigating the in vivo and in vitro activity of the cells associated with the loosening process in seven dogs. The cells were isolated and maintained in culture for sufficient periods of time so that their biologic activity could be studied as well as the effect of different agents added to the cells in vivo or in vitro. The biologic response as determined by interleukin-1 and prostaglandin E2 activity paralleled the roentgenographic appearance of loosening and the technetium images and observations made at the time of revision surgery. The correlation between clinical, roentgenographic, histologic, and biochemical loosening indicates that the canine model is suitable for investigating the mechanisms of prosthetic failure. A canine model permits the study of possible nonsurgical therapeutic interventions with the ultimate hope of stopping or slowing the loosening process.

Here is the redacted text:
...investigate the effects of an [REDACTED] agent on tissue changes around loose prostheses. The aseptically loosened prosthesis provided a means for investigating the in vivo and in vitro activity of the cells associated with the [REDACTED] process in seven dogs...
Note: I replaced "general pathological conditions" with "[REDACTED]" since it was requested to replace the specific phrase related to the reason why general pathological conditions are being investigated instead of neoplasms. Let me know if you have any further requests!

Figure 37: Example of prompting the `meta-llama/Meta-Llama-3-8B-Instruct` model on the medical dataset to output a redacted text which does not allow to explain both labels. This is an extension of the method from [39]. The model masks tokens with [REDACTED] that it considers crucial for answering the contrastive question. This redaction is applied additional $k = 5$ times (6 in total) for quantifiable results. We can notice that models are not necessarily robust at following prompts (e.g., the model did not give only the redacted text)

What is the class of the following text? The text can contain redacted words marked with [REDACTED]. Answer only general pathological conditions, neoplasms, or unsure. Do not explain the answer.

Here is the redacted text:
...investigate the effects of an [REDACTED] agent on tissue changes around loose prostheses. The aseptically loosened prosthesis provided a means for investigating the in vivo and in vitro activity of the cells associated with the [REDACTED] process in seven dogs...
Note: I replaced "general pathological conditions" with "[REDACTED]" since it was requested to replace the specific phrase related to the reason why general pathological conditions are being investigated instead of neoplasms. Let me know if you have any further requests!

General pathological conditions

Figure 38: Example of prompting the `meta-llama/Meta-Llama-3-8B-Instruct` model on the movie reviews dataset to predict the label on the redacted text. If the model predicts "unsure", then it is self-consistent. In any other case, it is not self-consistent as it claimed to redact words necessary for it to predict the label (answer the contrastive question). In this case, the model is not self-consistent, as it replied "general pathological conditions".

D.2.3 Explanation Generation

For the following text, why is it general pathological conditions?

Tissue changes around loose prostheses. A canine model to investigate the effects of an antiinflammatory agent. The aseptically loosened prosthesis provided a means for investigating the in vivo and in vitro activity of the cells associated with the loosening process in seven dogs. The cells were isolated and maintained in culture for sufficient periods of time so that their biologic activity could be studied as well as the effect of different agents added to the cells in vivo or in vitro. The biologic response as determined by interleukin-1 and prostaglandin E2 activity paralleled the roentgenographic appearance of loosening and the technetium images and observations made at the time of revision surgery. The correlation between clinical, roentgenographic, histologic, and biochemical loosening indicates that the canine model is suitable for investigating the mechanisms of prosthetic failure. A canine model permits the study of possible nonsurgical therapeutic interventions with the ultimate hope of stopping or slowing the loosening process.

The text describes research on tissue changes surrounding loose prostheses in canines (dogs). This falls under the category of ""general pathological conditions"" because:
* Loose prostheses are considered a type of abnormal condition or disease state. * The research aims to understand the underlying biological processes involved in this condition, including cellular responses and inflammatory reactions. * The investigation involves studying various aspects of the condition, such as radiological appearances, surgical findings, and biochemical markers.
In summary, the text focuses on understanding the pathophysiology of a specific disease-like condition (loose prostheses) using animal models, which makes it relevant to the field of general pathology.

Figure 39: Example of prompting the `meta-llama/Meta-Llama-3-8B-Instruct` model on the medical dataset to generate non-contrastive self-explanation NCE for the fact label.

For the following text, why is it NOT neoplasms?

Tissue changes around loose prostheses. A canine model to investigate the effects of an antiinflammatory agent. The aseptically loosened prosthesis provided a means for investigating the in vivo and in vitro activity of the cells associated with the loosening process in seven dogs. The cells were isolated and maintained in culture for sufficient periods of time so that their biologic activity could be studied as well as the effect of different agents added to the cells in vivo or in vitro. The biologic response as determined by interleukin-1 and prostaglandin E2 activity paralleled the roentgenographic appearance of loosening and the technetium images and observations made at the time of revision surgery. The correlation between clinical, roentgenographic, histologic, and biochemical loosening indicates that the canine model is suitable for investigating the mechanisms of prosthetic failure. A canine model permits the study of possible nonsurgical therapeutic interventions with the ultimate hope of stopping or slowing the loosening process.

The reason this tissue change is not considered a neoplasm (abnormal growth of cells) is because:
* There is no mention of abnormal cell proliferation or uncontrolled growth. * The ""cells"" being referred to are likely macrophages, fibroblasts, or other types of normal cells responding to inflammation and foreign body reaction caused by the loosened prosthesis. * The authors are studying the biological responses of these cells to various stimuli, such as inflammatory mediators, rather than describing any unusual or malignant cellular behavior.
In contrast, neoplasms typically involve aberrant cell growth patterns, which can lead to tumor formation. In this case, the tissue changes observed are related to a non-malignant, chronic inflammatory response to a foreign object (the loosened prosthesis).

Figure 40: Example of prompting the `meta-llama/Meta-Llama-3-8B-Instruct` model on the medical dataset to generate non-contrastive self-explanation NCE for the foil label.

For the following text, why is it general pathological conditions and not neoplasms?

Tissue changes around loose prostheses. A canine model to investigate the effects of an antiinflammatory agent. The aseptically loosened prosthesis provided a means for investigating the in vivo and in vitro activity of the cells associated with the loosening process in seven dogs. The cells were isolated and maintained in culture for sufficient periods of time so that their biologic activity could be studied as well as the effect of different agents added to the cells in vivo or in vitro. The biologic response as determined by interleukin-1 and prostaglandin E2 activity paralleled the roentgenographic appearance of loosening and the technetium images and observations made at the time of revision surgery. The correlation between clinical, roentgenographic, histologic, and biochemical loosening indicates that the canine model is suitable for investigating the mechanisms of prosthetic failure. A canine model permits the study of possible nonsurgical therapeutic interventions with the ultimate hope of stopping or slowing the loosening process.

The reason why ""general pathological conditions"" are mentioned instead of ""neoplasms"" (which refers specifically to abnormal cell growths) is likely because the study focuses on non-cancerous tissue changes caused by the presence of a loose dental prosthesis.
In this context, ""pathological conditions"" refer to any abnormal or diseased state of tissues, which can include inflammation, degeneration, or other types of damage caused by various factors such as mechanical stress, infection, or trauma. In contrast, neoplasms typically imply the development of cancer cells, which is not relevant to this specific research question.
By studying the general pathological conditions surrounding the loose prosthesis, researchers aim to understand how these abnormalities develop and progress over time, and how they respond to treatment interventions like anti-inflammatory agents. This knowledge may ultimately help improve our understanding of prosthetic failure mechanisms and inform strategies for preventing or treating related complications."

Figure 41: Prompting `meta-llama/Meta-Llama-3-8B-Instruct` model on the medical dataset to generate a modified contrastive self-explanation ("Why A and not B?"). The explanation is later evaluated on generality with the original CE (Figure 36).