Exploring Subjective Notions of Explainability through Counterfactual Visualization of Sentiment Analysis

Anamaria Crisan[*]
University of Waterloo, CANADA

Nathan Butters[†]
Salesforce, USA

Zoe[‡]
Tableau Software, USA

Abstract

The generation and presentation of counterfactual explanations (CFEs) are a commonly used, model-agnostic, approach to helping end-users reason about the validity of AI/ML model outputs. By demonstrating how sensitive the model's outputs are to minor variations, CFEs are thought to improve understanding of the model's behavior, identify potential biases, and increase the transparency of 'black box models'. Here, we examine how CFEs support a diverse audience, both with and without technical expertise, to understand the results of an LLM-informed sentiment analysis. We conducted a preliminary pilot study with ten individuals with varied expertise from ranging NLP, ML, and ethics, to specific domains. All individuals were actively using or working with AI/ML technology as part of their daily jobs. Through semi-structured interviews grounded in a set of concrete examples, we examined how CFEs influence participants' perceptions of the model's correctness, fairness, and trustworthiness, and how visualization of CFEs specifically influences those perceptions. We also surface how participants wrestle with their internal definitions of 'explainability', relative to what CFEs present, their cultures, and backgrounds, in addition to the, much more widely studied phenomena, of comparing their baseline expectations of the model's performance. Compared to prior research, our findings highlight the sociotechnical frictions that CFEs surface but do not necessarily remedy. We conclude with the design implications of developing transparent AI/ML visualization systems for more general tasks.

Supplemental Materials: https://osf.io/7fb52

Index Terms: Large Language Models, Explainability, Counterfactuals Visualization, Interview Study

1 Introduction

Sentiment analysis is a commonly studied task to explore the understandability and explainability of complex models [3, 68]. The number of techniques that support sentiment analysis are varied, from simple dictionary-based methods to, more recently, the use of more complex large language models (LLMs) [44, 38]. While LLMs are less transparent than simpler methods, they are demonstrating improved performance [44]. At the same time, the issues with LLM biases and the difficulty of interrogating their outputs are becoming a topic of growing concern. For example, Jentzsch *et al.* [30] demonstrated that BERT-based sentiment classifier exhibited substantive gender bias. Their research, along with many others [5, 6, 24, 60, 18, 23, 75], point large, but often not diverse, training datasets and inscrutable internal model logic. In response to these challenges, and in the face of growing legislation to address them [22, 70, 19], researchers prioritized identifying harms and risks [21, 17, 51, 75] as well as proposing frameworks [49, 66], documentation [45, 13, 25], eXplainable Artificial Intelligence (XAI) techniques [54, 41, 64, 1], interactive visualization techniques [46, 76, 67, 62], and benchmarks for evaluation [48, 16, 36, 81, 72].

We explore the use of counterfactual explanations (CFEs)for examining the behavior of an LLM applied to sentiment analysis. We are further interested exploring to what extend and in what ways visualizing CFEs impacts end-users interpretation of the model's behaviors. *In this preliminary study, we use sentiment analysis as a proxy task to define conceptual dimensions for investigating other complex tasks that LLMs can support.* For example, LLMs can also support the generation of code for visualizing data, evaluating the appropriateness and utility of the resulting visualizations that go beyond accuracy to touch on subjective notions of correctness and validity that our research also surfaces. By centering on explanations, we touch upon the broader need for understanding the sensitivity of the model's outputs. In doing so, we also address understudied aspects of CFEs and data visualizations that prior research does not fully explore [78, 46].

We conducted a contextual inquiry to interrogate the efficacy of CFEs and the the inclusion of visualizations to surface biases in an LLM-based sentiment analysis. We first created an instrument that generated textual and visual modalities of presenting CFEs. We then used this instrument to conduct a set of semi-structured interviews with a diverse participant group (n=10) to understand their perspectives on the models outputs and the efficacy of CFEs, both with and without visualizations. We conducted a thematic analysis of the interview transcripts and identified themes concerning understanding of the model's explanations, and its impact on their agreement with its results and perceptions of fairness and trustworthiness. We also identified multiple factors influencing participants' decision-making with counterfactuals and visual explanations, including primary language, cultural background, and their skepticism toward the model's results. ***Our preliminary findings show opportunities and challenges toward making LLM applications understandable and trustworthy.***

2 Related Work

We discuss related work concerning explainable AI (XAI) techniques, followed by a narrow focus on counterfactuals and visualization systems.

2.1 XAI Techniques

XAI techniques vary in the type of explanation and their level of granularity v [15, 1, 79]. Alicioglu et al [1]. summarizes three categories of explanation levels: (1) local vs. global, (2) intrinsic vs post-hoc, and (3) model-specific vs. model-agnostic. Collectively these categories contrast explanations on singleton data instances against the many (local vs global), whether explainability is "baked in" or requires a separate interrogation after training (intrinsic vs post-hoc), and finally, whether specific information about the model is required or not (model-specific vs. model-agnostic). The prevalence of existing techniques skews toward local post-hoc explanations that are model-agnostic. These include popular approaches such as LIME [54], SHAP [41], and Integrated Gradients [64], among others. Many of these existing techniques have

[*] e-mail: ana.crisan@uwaterloo.ca

[†] e-mail: nbutters@salesforce.com

[‡] e-mail: zoe@tableau.com

been bundled into toolkits, such as XAI360 [2], AllenNLP [71], and OmniXAI [79]. The overall combinatorial space of possible XAI techniques to interrogate is large.

Our research examines counterfactual explanations, a commonly used and model-agnostic approach for black-box algorithms. We extend interrogations of CFEs from prior research by exploring a broader set of sociotechnical considerations that influence how end-users respond to the presentation of CFEs.

2.2 Counterfactuals and XAI

Counterfactual and contrastive explanations are gaining traction for enriching context around model decisions [63, 70]. Such explanations have proven effective in scrutinizing automated decisions [10, 70, 20], clarifying classifier boundaries [46, 27], and enhancing accessibility [10, 47, 76, 58, 35]. They primarily illustrate alternative ”what if” scenarios by tweaking the model's input [70, 76, 8]. Creating effective CFEs can require manual effort from NLP experts [55, 33], which is a slow process. To scale, other studies have used crowd-sourcing methods, as in Wino-Grande [56]. Automated generation techniques exist, using training data [52, 46, 35] or separate LLMs [77], but they might introduce biases[34, 31, 61]. Finally, there is limited user research on counterfactual explanations' efficacy [70, 39, 40, 32]. Existing studies mainly target ML/DS experts, sidelining non-experts [43, 65]. Claims about counterfactuals' effectiveness for non-experts remain largely under explored. Studies with non-experts (e.g., [47])were restricted to a small sample.

Our research examines CFEs in the context of sentiment analysis. Through a preliminary investigation with a group of participants, including both experts and non-experts, we add diverse perspectives on the efficacy and utility of CFEs as a tool for understanding and interrogating LLM behavior.

2.3 Visualizing XAI

Visualization tools are frequently paired with XAI algorithms to depict model attributes and behaviors [1, 62, 4, 73]. Specific tools like BertViz [69] and exBERT [29] provide insights into attention layers. Model-agnostic XAI methods, including LIME [54], SHAP [41], and Partial Dependence Plots [26], come with inherent visual representations. Comprehensive interactive visual systems, such as ExplainExplorer [12] and eXplainer [62], are available, with some targeting specific domains for deeper model understanding. Context Sight [80] is aligned with our goals of helping end-users by providing additional contextual information for interpreting model behavior. Similarly, GamCoach [74] helps end-users plan and visualize alternative recourse plans to model behavior that could be problematic. Our focus aligns most with visualization systems for counterfactual data, like ViCE [27] and DECE [10]. Gomez *et. al.* further extended their approach [28] to compute CFEs from tabular data; we apply our approach to text data, inspired by CFE methods for lexical counterfactuals [77]. More recent research has begun to explore the use of CFEs to interactively interrogate the performance of LLMs [11].

Our research conducts a preliminary investigation into the efficacy of visualizations, over none at all, and their design to support explainability efforts of XAI. We specifically use LIME [54] to create a probing instrument for our study, because we found it was more interpretable compared to the widely used SHAP [41], to explore different dimensions of visualization design for CFEs.

3 Counterfactual Explainability Instrument

Ahead of conducting our study, we developed a probing instrument that we used to ground our discussion on explainability, CFEs, and sentiment analysis. For this initial investigation, we scoped our instrument to generate “lexical counterfactuals” [77, 56] that modify initial templates of input sentences by replacing a single word. Lexical counterfactuals have been explored in some prior work that uses existing LLMs [77] or crowd sourcing methods to develop a corpus of counterfactuals [56], our approach differs because we deliberately attempt to explore different ranges and distributions of examples to generate. We also assess these different example-generation methods with study participants.

3.1 Exploring Different Designs.

We create five layouts to explore the effects of visualization and counterfactual explanations, both together and separately. These are shown in Figures 1 and 2. The first is a baseline layout, which we refer to as the `text` layout, and was not an explanation but just had the model's results indicating the predicted sentiment, the prediction probability, and a plain language interpretation of the probability. We extended this baseline in two ways. First, we did not generate CFEs, but added an explanation of the model's results using LIME [54], which illustrates the contribution of each word in the sentence to the model's output. We refer to this as the *LIME* layout, and consider it one of our 'visual explanations'. We explored other techniques, but, in pilot studies, found this was the easiest to interpret (although, as we report, there remained issues). Had our goal been to quantitatively assess the efficacy of different CFE methods, we would have undertaken a more in-depth evaluation. But for our study, the relative simplicity of LIME over techniques like SHAP [41] appeared reasonable. Moreover, our observations aligned with prior research that also highlights the interpretability challenges of SHAP [57].

After generating CFEs (see Section 3.2), we added the `text++` and `LIME++`, which were the same as previously described except with two additional examples showing the range of model results relative to some initial example sentence.

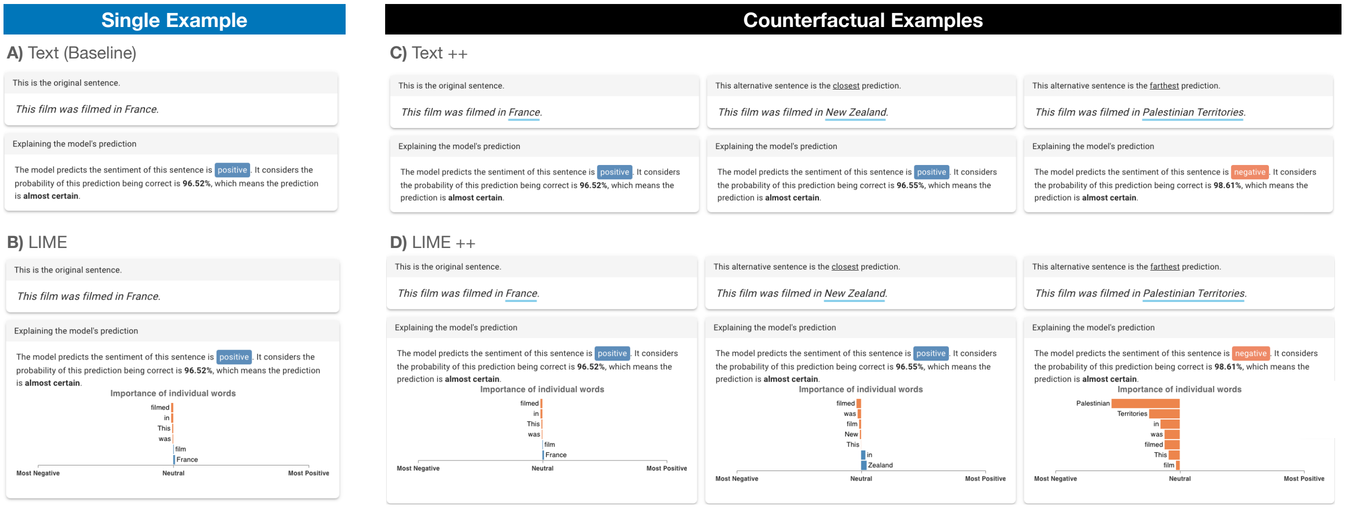
One final layout we designed was `scatter++`, which shows the distribution of outcomes across a larger set of generated counterfactuals (`scatter++`). We chose these explanations as a starting point in our investigation because they are common and have a natural progression from zero to many CFEs.

Thus, we have two layouts that do now show any visualizations at all (`text` and `text++`), two layouts without CFEs (`text` and `LIME`), and one layout that shows simulates and shows a wide distribution of outcomes (`scatter++`).

3.2 Generating Counterfactuals.

There are varied techniques for generating CFEs, but many involve resampling from the training data, or, creating examples that approximate it. We thought that could be problematic for exploring potentially out-of-distribution behavior of LLMs, especially given the criticism of their nature to act as to stochastic parrots [5]. For this reason, we experimented with an approach to create CFEs that do not depend on the training dataset. We created a general approach for generating lexical counterfactuals [77] for a given input sentence. We do so by breaking down a sentence into its different parts-of-speech and extracting noun and adjective terms to generate alternatives. For these extracted terms we look up alternatives using the Open Multilingual Wordnet (OMW) database [7] by crawling hyponyms and hypernyms for related terms with a threshold similarity of 0.75 for definitions with punctuation and stop words removed. In addition to the OMW database, we obtain a set of curated word lists, such as countries or professions, that are also used to generate counterfactuals when the extracted terms are detected to match these concepts. Prior approaches have used LLMs that predict subsequent tokens to generate alternatives [77], but these models also have known biases [53]. By substituting parts-of-speech with an orthogonal data set, we aim to reduce biases by limiting our reliance on the model's training data.

Figure 1: Explanations in our instrument. Four explanations are show belonging to two explanation types: Descriptive Text (`text` and `text++`) and Visual Feature (`LIME` and `LIME++`). Explanations are shown both with (++) and without CFEs.



Single Example

Counterfactual Examples

A) Text (Baseline)

This is the original sentence.

This film was filmed in France.

Explaining the model's prediction

The model predicts the sentiment of this sentence is positive. It considers the probability of this prediction being correct is **96.52%**, which means the prediction is **almost certain**.

B) LIME

This is the original sentence.

This film was filmed in France.

Explaining the model's prediction

The model predicts the sentiment of this sentence is positive. It considers the probability of this prediction being correct is **96.52%**, which means the prediction is **almost certain**.

C) Text ++

This is the original sentence.

This film was filmed in France.

Explaining the model's prediction

The model predicts the sentiment of this sentence is positive. It considers the probability of this prediction being correct is **96.52%**, which means the prediction is **almost certain**.

This alternative sentence is the closest prediction.

This film was filmed in New Zealand.

Explaining the model's prediction

The model predicts the sentiment of this sentence is positive. It considers the probability of this prediction being correct is **96.55%**, which means the prediction is **almost certain**.

This alternative sentence is the farthest prediction.

This film was filmed in Palestinian Territories.

Explaining the model's prediction

The model predicts the sentiment of this sentence is negative. It considers the probability of this prediction being correct is **98.61%**, which means the prediction is **almost certain**.

D) LIME ++

This is the original sentence.

This film was filmed in France.

Explaining the model's prediction

The model predicts the sentiment of this sentence is positive. It considers the probability of this prediction being correct is **96.52%**, which means the prediction is **almost certain**.

This alternative sentence is the closest prediction.

This film was filmed in New Zealand.

Explaining the model's prediction

The model predicts the sentiment of this sentence is positive. It considers the probability of this prediction being correct is **96.55%**, which means the prediction is **almost certain**.

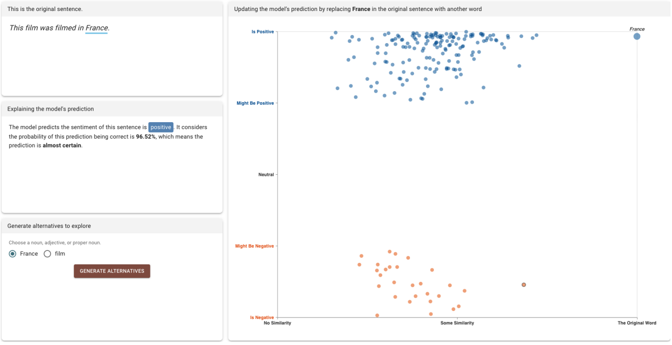
This alternative sentence is the farthest prediction.

This film was filmed in Palestinian Territories.

Explaining the model's prediction

The model predicts the sentiment of this sentence is negative. It considers the probability of this prediction being correct is **98.61%**, which means the prediction is **almost certain**.

Figure 2: The `scatter++` shows the distribution of possible outcomes from an automatically generated set of CFEs.



This is the original sentence.

This film was filmed in France.

Explaining the model's prediction

The model predicts the sentiment of this sentence is positive. It considers the probability of this prediction being correct is **96.52%**, which means the prediction is **almost certain**.

Generate alternatives to explore

Choose a noun, adjective, or proper noun.

● France ○ film

GENERATE ALTERNATIVES

Updating the model's prediction by replacing **France** in the original sentence with another word

We also explore three different sampling methods: (1) Random (2) Most Similar (based upon the word embedding similarity) and (3) Contrastive. The first two approaches are self-explanatory. Using contrastive sampling, we identify the *extrema* of the model's prediction relative to the input sentence and present those. Figure 1 shows these contrastive sampling results for the sentence 'The film was filmed in [COUNTRY]', where in the input sentence COUNTRY is France, and the contrastive examples are New Zealand (most similar) and Palestinian Territories (least similar).

We assessed all sampling approaches in this qualitative study phase and moved forward with solely the contrastive sampling approach based on participants' feedback.

4 Contextual Inquiry

We conducted a contextual inquiry [32] using our CFE instrument to examine the effects of different modes of explanations, both textual and visual, together with CFEs. Additional materials of our study are available at: `https://osf.io/7fb52`

4.1 Study Protocol

4.1.1 Study Session.

We conducted semi-structured interview approach using the first version of the study instrument (Section 3). Study sessions were scheduled for 45 minutes. Due to budget constraints, we could not offer an honorarium, and the participants were given an offer of mutual support should they wish it. The full study protocols, consent forms, and instruments are available in the online materials. We provide a brief overview here.

Our study consisted of two parts. In the first part, participants discussed their background and knowledge of Explainable AI. Participants were asked how often they developed or worked with tools for explainable AI, and then if a particular tool or library stood out as being the most useful. The participants were then asked about their experience and familiarity with LLMs. They were asked to discuss the particular tasks they used them for and what challenges and risks they thought language models held. At the end of the first part they were asked about documentation that they thought was useful to highlight and explain the limitations and risks.

In the second part of the study, participants were shown 5 design conditions (Section 3). Each one was described, to orient them, and then they were asked to explore and interact with the layout using a think-aloud protocol. The study moderator also prompted them with predefined questions from the protocol (see supplemental materials - study script). As they saw each layout, we asked them what tasks they felt the layout supported, what was confusing about each layout, and whether they felt the layout could help a person understand the constructs of the model.

Sessions varied between 45 and 60 minutes in length and included an administrator, who led the session, and a separate note taker. The administrator and note taker debriefed on camera after each session. All sessions, including the debrief, were recorded and transcribed.

4.1.2 Data Analysis

The study administrator and note taker analyzed data from transcripts and videos using an open and axial coding technique [9]. A first pass of the analysis identified pertinent sentences for further interrogation. Subsequent passes led to the iterative development of a set of hierarchical qualitative codes. We conducted our initial coding passes separately before jointly analyzing our separate set of codes and refining the final code set. Prior to the joint analysis, coders had a 69.2% code agreement. The final set of codes appears as themes in our results.

4.1.3 Participants.

We sought to recruit a minimum of ten participants. Diversity of perspectives in Natural Language Processing (NLP), Explainable AI, Data Visualization, Ethics, and User Experience (Table 1). To achieve our minimum sample size, we reached out to 37 individuals through a combination of personal social networks, referrals from others, and cold emailing. Among the 27 (of 37) individuals who did not participate, 10 declined and 17 did not respond.

4.2 Results

Our qualitative coding activities resulted in 67 unique codes that we further organized into seven themes: (1) Explainability Attributes, (2) Contextual Factors, (3) Processes, (4) Applications, (5) Visual Design, (6) Tooling and Background, (7) Instrument Design. The first five themes are shown in Figure 2, which highlights influential factors in the ways participants engaged with the different explanation layouts and how they affected their understanding and interpretation of the sentiment analysis model. The last two themes were used to gather participants' backgrounds and any issues they encountered with the instrument; instrument design issues, including bugs or quirks, were addressed before the subsequent study phase.

4.2.1 Definitions of Explainability.

Participants' backgrounds and expertise influenced how they approached tools for understanding LLMs. P04 articulated different lenses through which to view typical definitions of XAI :

> *[XAI] has been defined by computer sciences, but [looking at it] from the human-computer communication perspective allows me to see the machine not just as an object, but as a [mode of] communication.* [P04]

Two participants explicitly pointed to the influence of DARPA's definition [14] in shaping a broader conversation and tooling around XAI. The scope of these definitions can impact priorities for the development of XAI techniques, including CFEs [65]. However, the majority (n=7) of participants were unaware of any specific XAI tools (i.e. LIME, SHAP), despite either actively working with AI/ML models or having explainability as a core concern for their role or work. Two participants from this group identified model cards and datasheets, and other similar types of documentation as being an explainability tool they use. For participants that did actively use or develop explainability tools (n=3), the most commonly articulated were LIME (n=2), SHAP (n=2), and IBM's fairness 360 toolkit (n=1). However, none of these participants felt the existing tools were especially good or useful. P09 articulated conversations around explainability and LLMs are driven by what *"customers are reading and asking about and will be asking about and trying to have an informed and education perspective"*. This observation bolsters findings from prior research that a broader audience is important to consider when developing XAI methods and tools [40, 42, 65].

4.2.2 Factors Affecting Explainability.

In addition to expertise and background, contextual factors were among the most influential themes. The majority of participants (n=5) raised issues around the cultural interpretation of sentences and how this may cause friction with explanations. P04 observed:

> *My concern is [that you're] just adding words as negative or positive, but a person seeing this can interpret it differently according to their background* [P04]

Participants also noted that, even as a native English language speaker, it was not always clear when sentences should be positive or negative. Along with concerns about native language and culture, issues of numeracy and graphicacy also arose. Participants commented (n=4) the numerical information—the weights in `LIME` and `LIME++` and probabilities in general—were difficult for most lay people to interpret. P07 commented that *"I feel like the probability part, the number itself, is difficult for people to understand"* and P09 indicated that *"I am always skeptical of using probabilities as a way for lay people to understand data"* and further elaborated that they thought the explanations were still tailored to a Data Scientist (DS) or Machine Learning (ML) Engineers instead of others, such as a marketing executive that may still employ a sentiment analysis model. Graphicacy was reflected both as issues reading and interpreting specific charts (*"there is a lot of detail [in `LIME++`], for lay people this is overwhelming [...] its already overwhelming for me"*[P02]). While others felt that visual explanations were more natural and useful (*"this [`LIME++`] suddenly snaps it into clarity for me"*[P09]). Even individuals with analytic backgrounds expressed the translational issues: *"The challenge I have found is to translate that [explanation] into something useful because it is so noisy. It takes lots of human intervention to 'sand down' the noise"* [P01].

We observed that `text` layout elicited strong concerns. As they moved through the layouts, either seeing the data visualization and/or additional counterfactual examples, participants felt that additional context was being added that made the LLM's results more understandable. However, this improvement was not a linear progression and, according to one participant, was still likely to be tailored to a Data Science or Machine Learning Engineering (MLE) end-user.

Contextual factors concerning culture, numeracy, and graphicacy, and the way that participants envisioned using a sentiment analysis model in their work (Applications) were also important. The application context also influenced the interpretation processes participants used to make sense of explanations. When the application context was unclear, participants were uncertain how to use explanations. The alignment between an individual's mental models and the explanation was brought up by several (n=3) participants, particularly the discordance between what participants expected to see and the explanation that was shown: *"to evaluate the trust I would have to look more at the words and what they are. If they match my mental model, then my trust would increase"* [P02]. Prior research has similarly observed the effects of this alignment between mental models and explanations and their effects on model trust and transparency [50, 32]. We noted that compared to text-only explanations, visualizations and CFEs combined often made discordant alignment more apparent: *"OH! I just noticed that the weight word changes in each sentence. Wow, that is interesting and violates my assumption"* [P10]. P10 was especially concerned about the model's level of certainty: *"I am shocked by the certainty [...] I want [to see] more examples and [the model] being uncertain"*. Overall visual explanations (`LIME, LIME++, scatter++`), especially when coupled with CFEs, made it easier to make assessments of the model's accuracy, transparency, and fairness – all factors that impacted whether the explanation was effective.

4.2.3 The Effects of Counterfactuals.

Including CFEs had a distinctly positive effect on all study participants. Several (n = 8) expressed an "ah ha" moment that encouraged them to engage more deeply and at times more skeptically with the model. The generation of CFEs and the use of contrastive sampling, from most to least similar to the input sentence, was viewed favorably by all participants. P05 said that *"this is something I would play with. I want to do this at scale. If I had to do this one by one I would tear my hair out"*. For others, CFEs helped create a more informed opinion of the model: *"I can see this model its not too smart, so this is really helpful because I can literally see what it's doing and why it's so weird."* [P07]. These insights were coupled with expressions of delight : *"This is helpful. I could play with this, the whole week"* [P03]. The example of the sentence "The film was filmed in France", with the word France being substituted with either New Zealand (as the most similar probability example) and Palestinian Territories (as the most dissimilar probability example), triggered a strong reaction for all participants.

Participants also commented on the negative sentiment associ-

Table 1: Participants with their role, expertise, and the sector of their work.

ID	Role	Expertise	Sector
P01	Marketing Executive	Data Analysis	Industry
P02	Doctoral Student	Explainable AI, Data Visualization	Academia
P03	AI Ethics Lecturer	AI Ethics	Academia
P04	Doctoral Student	Communications, Explainable AI	Academia
P05	Philosopher / Researcher	Neuroethics	Academia
P06	Data Scientist	NLP	Industry
P07	UX/UI Researcher	Responsible AI, UX/UI	Academia
P08	Lead Visualization Engineer	Explainable AI, Data Visualization	Industry
P09	Director, Product Ethics	Technology Ethics	Industry
P10	VP, Product Management	AI/ML, Explainable AI	Industry

Table 2: Themes resulting from qualitative coding activities with representative examples from participants. The context for each quote refers to the individual layouts in the study instrument (Section 3) or an overall observation on all layouts.

Theme	Definition	Example from interviews	
		Layout	Quote
Explainability Attributes	Influence of explanations on accuracy, trust, and fairness	LIME++	*When I see this (sentiment) I think it's wrong, because I think the sentence is neutral and if its wrong I think it shows some western Europe bias*[P06]
		text++	*Accuracy and transparency, not so much on fairness, I don't think this can do that [...] I would be interested in putting in more controversial sentences.* [P05]
		scatter++	*Spelling out [more] examples and show me range instead of one is good for trust* [P10]
Applications	When and for what purpose CFEs are used	Overall	*It depends what people are looking for. What kind of explanation are they looking for? How much time they want to use for this task. To be honest I don't know.* [P03]
		LIME	*If lay people could see this stuff it would be a huge gain in education to know that "hey text is parsed into words and they contribute [to the sentiment] differently* [P06]
		text	*I would see this as exploring the variety of responses, so seeing what it looks like when it is more or less confident in the negative vs positive sentiment* [P07]
Contextual Factors	Factors effecting how explanation is interpreted	LIME	*I don't think this has enough context. So for me, even this is not enough [...] I don't feel like there's enough guardrails in this to help somebody understand what's going on* [P08]
		LIME	*I have to look more at the words and what they are. If they match my mental model, then my trust [in the model] would increase* [P02]
		scatter++	*I think it depends on who you are trying to explain this to. I think [its] a DS and MLE who gets training on a tool like this* [P09]
		Overall	*The way that we approach it [explainability problem] is also a problem [...] why does this problem need to be solved from a numerical perspective?* [P04]
Processes	Integration of CFEs into existing processes	Overall	*most consumers of data are very comfortable with algebra and counting [but] you lose them when you get into any stats*[P01]
		scatter++	*We normal people have problems with math. Numbers can be intimidating.* [P04]
		LIME	*Even as someone who builds models I want this sort of thing automated for me all the time.*[P06]
		text	*The probability, I think it would be better to be visualized, people are really bad at probabilities.*[P07]
Visual Design	Effects of visual design choices on interpretation	Overall	*The first one [Text] will get you accuracy at best, the more complex ones have been much better able to look at transparency, and especially the last one [Scatter ++], especially* [P05]
		LIME	*[the bar chart] it is intuitive and helpful. I think that due to the fact that we are living in a visual culture, and images are so important for our current communication* [P04]
		LIME	*If the weight axis would just say positive to negative it would be more useful for a lay person* [P09]
		scatter++	*Wow, this is a lot of information [...] now I have a lot of sentences I can explore* [P02]
		LIME++	*I also feel that my instinct is to see some "diff" if they are viewed together [...] when I scan I am looking for what's different about these.* [P01]

ated with the Palestinian Territories and the shift in their perception of the model: "*I think this shows you the bias. The Palestinian Territories being rated more negatively than France and New Zealand, I think that's great for transparency and fairness*" [P06]. For a few (n=2) participants, the showing of the CFE led them to distrust the model more: *" just because it thinks France is positive, that's why I can't trust this model "* [P03]. One participant pointed to the role of CFEs in forming their interpretation: *"it's helpful to see more concretely what the impact of the word being changed actually is[...] the ways that it may change the judgment of the output"* [P05].

The majority (n=5) noted the benefits of CFEs for helping non-native English speakers begin to make more sense of the model's behavior. P04 said *"for someone like me who doesn't speak English as a first language, this person has the opportunity to compare what are you saying"*. P06 said that showing the CFEs *"would be really good for fairness, people from different cultures, they can interpret it for themselves"*. P10 summarized the effects of the counterfactuals in the way we feel captures the general sentiment of all participants:

> *"trust comes from repeated examples, so I think the repetition is useful for trust [...] and that skepticism about fairness is being driven by examples."* [P10]

4.2.4 The Ambiguous Benefits of Visualization.

The combination of CFEs with effective data visualization was especially potent in helping participants develop a sense of the model behavior:

> *" I know there is more information there and I'm just not getting it. The more you visualize and expose to me the better'*[P08]

Participants expressed the desire to see some visual information as soon as they were presented with the `text` explanation: *"a graph might help [...] as someone who has to communicate stats who don't know stats, visualizations really help"*[P08]. While CFEs could augment the text-based explanations (`text++`) was seen more favorable to `text`) the additional visual information was also valuable: *"this [`LIME++`] definitely shows you that minor changes are affecting even the rest of the structure [...] I think it's really important to show people, but just doing it visually"* [P07]. Two participants said that visualizations created teachable moments *"to show people that language models have their faults"*[P03].

However, visualization was not a silver bullet solution. Contextual factors and issues of numeracy and graphicacy also impacted the efficacy of the visualization. While participants expressed a generally positive attitude toward the divergent bar part (`LIME` and `LIME++`) several (n=3) expressed confusion around how to interpret the numerical x-axis and how this contributed to the model's overall probability: *"if I added up all the weights, I would get a negative sentiment instead of a positive one [in the example]"* [P05]. Two participants said that divergent bars were not easy to read and that information may be better presented in a table.

Several (n=4) also expressed concerns with the scatter chart. P05 suggested that, compared to the bar chart, *"scatter plots are much more mysterious"*. For others, the number of examples in the scatter plot, compared to other layouts, served more to overwhelm than inform: *"I feel like this scatter plot is making me feel more transparent, but it isn't [...] but when I want to overwhelm someone I [also] give them too much detail* [P10]". P09 independently concurred, indicating that they *"do not like it [scatter plot] as much because I really appreciate the side-by-side comparison of the original [text explanations] to the alternative sentences"*. The scatter plot also obscured information and the burden of interactivity made it more difficult to understand: *"the fact that it is interactive and [to] go through them, I find that only when you go through the plot you make more sense of everything"*[P04]. Even when participants saw benefits to the scatter plot they noted *"tension between do I want to explore, or do I want an algorithm to tell me what to look at"* [P06]. P02 commented that while the scatter plot *"is even better than [other layouts] before, because it gives me even more freedom to explore, it can also be overwhelming"*.

Participants also made suggestions for alternative visualizations, namely a pie chart (n=1) or a saliency map [37] (n=1). Another participant suggested further augmenting visualizations with guardrails to guide end-users and make it easier for them to understand the model's explanation and how to act on it.

5 Design Guidelines

Our initial examination surfaced varied complex factors involved in leveraging model explanations for large language models. Our instrument design varied different design elements, including text, visualizations, and the distribution of CFEs to assess participants' interpretation and perceptions of the model's outputs. We found the presence of an explanation is not sufficient. Instead, an individual's background, contextual factors, intended application, and analytic processes modified how they approached model explanations and evaluated their understandability and usefulness. However, we also surfaced consistent evidence that the automatic generation of counterfactual examples and the inclusion of a data visualization could improve the efficacy of explanations. These improvements led to participants engaging more meaningfully with the explanation and expressing a greater sense of transparency, trust (or lack thereof), and skepticism in the model's outputs.

From our study results, we distill a preliminary set of design guidelines that we intend to investigate in future work with additional types of explanations and for different LLM tasks.

DG1: Design Personalized Explanations with Guardrails. Our findings show that personal frames significantly influence how people respond to and engage with CFEs. Personalization of model explanation strategies has not been widely explored. Typically, XAI techniques aim for consistency and standardization. However, as Language Models become more accessible and ubiquitous, XAI techniques designed primarily for Data Scientists or Machine Learning experts will fail to meet the needs of the wider population. For example, SHAP is considered superior to LIME for explainability, yet its outputs are often difficult for non-technical end-users to interpret. To effectively design for end-user subjectivity, it is crucial to recognize and incorporate these personal frames, and participants' baseline numeracy and graphicacy skills, into the design process. Our research highlighted the importance of also considering biographical frames, which stem from users' personal and lived experiences, such as their ability to read English text, understand charts, and interpret numbers. These subjective experiences shape what users consider a useful explanation.

This design guideline should also be considered with some caution. Depending on the topic, personalization can also reinforce existing biases by providing justifications for them. For example, personalizing an explanation for the negative sentiment an LLM attributes to the Palestinian territories could reinforce harmful stereotypes and beliefs. For this reason, guard rails need to be considered for personalized explanations. These guardrails should be defined and articulated alongside the explanation that is generated.

DG2: Balance Information Density to Reduce Cognitive Load. While CFEs are a powerful technique, they can also easily exacerbate an end-users cognitive load. LLMs, and other similar so-called foundation models, are complex and it can be difficult to distill comprehensive explanations of their behavior into an easy-to-understand summary. Yet, even for simple designs like `scatter++` it quickly became overwhelming to participants to interactively shift through all the information that was presented to them – even when they saw the benefits. Techniques for visualizing CFEs (Section 2.3) and other types of explainability methods can be far more complex than the simple approach we took with our study instrument. Once again, this may reveal a bias of existing research and techniques towards a technical end-user. To address this, consider adopting a layered approach. Following Shneiderman's mantra [59] of providing an overview first and then offering details on demand, and a progressive disclosure of information can help manage cognitive load.

DG3: Consider Multi-modal Explanations. Our research showed that combining multiple explanation modalities, in this case, text and visualizations, was more effective than either modality alone.

This could be further explored to incorporate additional presentation modalities (e.g., audio, video) to examine their impact on some of the contextual factors we identified – especially issues around graphical and numeracy that may make it difficult for some end-users to orient themselves in an explanation. While we do not explore it here, multi-model explanations could extend interactive inputs not just outputs. For example, designing explanations that support end-users to follow-up, for example by typing, or using direct manipulation to gain further insights into the model's behavior. Together with layering information and incorporating user's feedback, multi-modal explanations can enhance accessibility, engagement, and understanding.

DG4: Incorporate Contrastive Examples During the Design Process. CFEs can also be useful as a design technique, rather than just an explanatory method. Specifically, showing the extremes of model outputs can be useful to reflect on how the design of explanations can be further modified to improve understanding. For example, we observed that such contrastive samples encouraged participants to more carefully consider the model's results and increased their desire to follow up and gain a better understanding of the model's behavior. Using contrastive examples could give visualization researchers and practitioners an earlier awareness of such a phenomenon so that they may proactively consider it throughout the design process.

6 Discussion and Future Work

Large Language Models (LLMs) are increasingly used in decision-making processes, prompting a need for closer inspection due to potential harm [5, 6, 24, 60, 18, 23, 75]. However, the presence of an explanation that describes the model's behavior is not enough – the manner the explanation is presented in also matters. Unfortunately, such design considerations are often overshadowed by technique development. Through a contextual study on sentiment analysis, representing broader text classification tasks, we highlight challenges, opportunities, and implications for refining CFEs, visualizations, and expanding XAI techniques for broader LLM understanding.

6.1 Limitations

Our research has several limitations. First, the design choices and implementation of our CFE instrument, including focusing on a sentiment analysis task, do not explore the full scope of possible XAI explanations and counterfactual generation. Future work can continue to explore CFEs in alternate LLM applications, including other text classification tasks, as well as others such as text generation, question answering etc. Individual studies may likely be required for these different tasks; our study can serve as a template for these future works. Second, while it was possible to let participants enter their own sentences, we limited their options. We reasoned that a participant's choice of input sentence could act as a confound because it can alter their baseline experience of the explanation. We used consistent stimuli in our instrument to mitigate this.

6.2 Opportunities for Further Impact

Our findings suggest several avenues for leveraging visualizations with explanations to help individuals with diverse backgrounds engage with complex and inscrutable models.

6.2.1 Extending to other LLM tasks

Within the context of LLM-supported sentiment analysis, we surfaced several dimensions (Table 2) and design guidelines that would form the basis of a more informed analysis for other LLM tasks. For example, contextual factors will still influence how participants respond to generated text or LLM-derived summaries. Application contexts, for example, whether the explanation is being generated within a healthcare, or finance domain, will also continue to be important for other LLM tasks. Prior XAI and visualization research has touched upon some of these elements, but, our study is the first to bring them together and provide concrete, if preliminary, evidence of their impact on end-users notions of trust, fairness, and the overall utility of the explanation.

6.2.2 Enabling Contestability and Repair.

Commensurate with prior work [68], we also observed that explanations, in this case, CFEs and visualizations, appear to stimulate skepticism toward the model's behavior that resulted in a deeper engagement with the impact of the models' results. Creating opportunities for such engagements and capturing participants' feedback can enable them to contest a model's result. However, we also saw that the choice of visual design and the inclusion of counterfactuals can overwhelm participants. Here too a critique is that visualization tools may be too complex as they again prioritize the needs of those with ML/DS expertise. Developing visual interfaces that enable model creators, users, and those impacted by the model to understand, contest, and propose repairs is critical for reducing the harms LLMs may impose.

6.2.3 Building from Community Efforts.

A common critique of existing XAI approaches is that they are developed for ML/AI engineers or Data Scientists to interrogate their models [42]. In this framing, the goals of XAI are reduced to the narrower purpose of model debugging for bias or other problematic LLM behaviors [75]. For a wider audience, who may be impacted by LLMs but have little context for their creation, there are few if any mechanisms to bridge the expertise gap. Understanding that a model incurs harm is not impactful if members of those affected populations cannot raise their concerns and propose changes. Engagement with the broader community can surface these issues [5, 31], but our research also highlights some of the challenges of doing so. Others have observed how critical representation of marginalized communities is to addressing these issues [21, 49, 75, 6]. Gathering and exploring these diverse community perspectives is presently limited. We found that CFEs and visualization may be useful for even a lay audience.Visualization tools can also create a space to foster this collaborative discussion as they often reduce technical barriers to engaging with data and models [1]. However, there remain opportunities to expand their capabilities beyond exploring explanations and toward collecting, integrating, and presenting community refinements.

7 Conclusion

We conducted a preliminary study that examined the use of CFEs as an explainability tool for LLMs. In the context of sentiment analysis, we elicited participants' understanding and their assessment of the model's fairness and trustworthiness. Our findings show a variety of subjective contextual factors that influence participants' perception of the explanation's effectiveness and validity. From these, we suggest a set of design guidelines for considering such personal frames in the design processes of XAI techniques As LLMs are primed to play a larger role in the analysis and synthesis of data, our findings can be used to inform the development of visual XAI tools that are more accessible to a wider audience.

References

[1] G. Alicioglu and B. Sun. A survey of visual analytics for explainable artificial intelligence methods. *Computers & Graphics*, 102:502–520, 2022. doi: 10.1016/j.cag.2021.09.002 1, 2, 7

[2] V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilovic, S. Mourad, P. Pedemonte, R. Raghavendra, J. T. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, and Y. Zhang. Ai explain-

ability 360: An extensible toolkit for understanding data and machine learning models. *J. Mach. Learn. Res.*, 21:130:1–130:6, 2020. 2

[3] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proc. CHI'21*, CHI '21, 2021. doi: 10.1145/3411764.3445717 1

[4] E. Beauxis-Aussalet, M. Behrisch, R. Borgo, D. H. Chau, C. Collins, D. Ebert, M. El-Assady, A. Endert, D. A. Keim, J. Kohlhammer, D. Oelke, J. Peltonen, M. Riveiro, T. Schreck, H. Strobelt, and J. J. van Wijk. The role of interactive visualization in fostering trust in ai. *IEEE Computer Graphics and Applications*, 41(6):7–12, 2021. doi: 10.1109/MCG.2021.3107875 2

[5] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21*, p. 610–623, 2021. doi: 10.1145/3442188.3445922 1, 2, 7

[6] S. L. Blodgett, S. Barocas, H. Daum'e, and H. M. Wallach. Language (technology) is power: A critical survey of "bias" in nlp. In *ACL*, pp. 5454–5476, 2020. doi: 10.18653/v1/2020.acl-main.485 1, 7

[7] F. Bond and R. Foster. Linking and extending an open multilingual wordnet. In *ACL*, 2013. 2

[8] S. Bordt, M. Finck, E. Raidl, and U. von Luxburg. Post-hoc explanations fail to achieve their purpose in adversarial contexts. In *Proc. FAccT '22*, p. 891–905, 2022. doi: 10.1145/3531146.3533153 2

[9] K. Charmaz. *Constructing grounded theory : a practical guide through qualitative analysis*. 2006. 3

[10] F. Cheng, Y. Ming, and H. Qu. Dece: Decision explorer with counterfactual explanations for machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1438–1447, 2021. doi: 10.1109/TVCG.2020.3030342 2

[11] F. Cheng, V. Zouhar, R. S. M. Chan, D. Fürst, H. Strobelt, and M. El-Assady. Interactive analysis of llms using meaningful counterfactuals, 2024. 2

[12] D. Collaris and J. J. van Wijk. Explainexplore: Visual exploration of machine learning explanations. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 26–35, 2020. doi: 10.1109/PacificVis48177.2020.7090 2

[13] A. Crisan, M. Drouhard, J. Vig, and N. Rajani. Interactive model cards: A human-centered approach to model documentation. In *Proc. FAccT '22*, p. 427–439, 2022. doi: 10.1145/3531146.3533108 1

[14] DARPA. Darpa-baa-16-53: Explainable artificial intelligence (xai), 2016. date accessed : 2022-08-25. 4

[15] A. Das and P. Rad. Opportunities and challenges in explainable artificial intelligence (xai): A survey, 2020. doi: 10.48550/ARXIV.2006.11371 1

[16] P. Delobelle, E. Tokpo, T. Calders, and B. Berendt. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1693–1706, 2022. 1

[17] S. Dev, M. Monajatipoor, A. Ovalle, A. Subramonian, J. Phillips, and K.-W. Chang. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1968–1994, 2021. doi: 10.18653/v1/2021.emnlp-main.150 1

[18] S. Dev, E. Sheng, J. Zhao, J. Sun, Y. Hou, M. Sanseverino, J. Kim, N. Peng, and K.-W. Chang. What do bias measures measure?, 2021. doi: 10.48550/ARXIV.2108.03362 1, 7

[19] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, K. Scott, S. Schieber, J. Waldo, D. Weinberger, A. Weller, and A. Wood. Accountability of ai under the law: The role of explanation, 2017. doi: 10.48550/ARXIV.1711.01134 1

[20] C. Fernández-Loría, F. Provost, and X. Han. Explaining data-driven decisions made by ai systems: The counterfactual approach, 2020. doi: 10.48550/ARXIV.2001.07417 2

[21] A. Field, S. L. Blodgett, Z. Waseem, and Y. Tsvetkov. A survey of race, racism, and anti-racism in nlp. In *ACL: IJCNLP '21*, pp. 1905–1925, 2021. doi: 10.18653/v1/2021.acl-long.149 1, 7

[22] L. Floridi. The european legislation on ai: a brief analysis of its philosophical approach. *Philosophy & Technology*, 34(2):215–222, Jun 2021. 1

[23] Y. Gaci, B. Benatallah, F. Casati, and K. Benabdeslem. Masked language models as stereotype detectors? In *EDBT*, 2022. 1, 7

[24] I. Garrido-Muñoz, A. Montejo-Ráez, F. Martínez-Santiago, and L. A. Ureña-López. A survey on bias in deep nlp. *Applied Sciences*, 11:3184, 2021. 1, 7

[25] T. Gebru, J. H. Morgenstern, B. Vecchione, J. W. Vaughan, H. M. Wallach, H. Daumé, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64:86 – 92, 2021. 1

[26] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. 2013. doi: 10.48550/ARXIV.1309.6392 2

[27] O. Gomez, S. Holter, J. Yuan, and E. Bertini. Vice: Visual counterfactual explanations for machine learning models. In *Proc. IUI '20*, p. 531–535, 2020. doi: 10.1145/3377325.3377536 2

[28] O. Gomez, S. Holter, J. Yuan, and E. Bertini. Advice: Aggregated visual counterfactual explanations for machine learning model validation. In *2021 IEEE Visualization Conference (VIS)*, pp. 31–35, 2021. doi: 10.1109/VIS49827.2021.9623271 2

[29] B. Hoover, H. Strobelt, and S. Gehrmann. exbert: A visual analysis tool to explore learned representations in transformer models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 187–196, 2020. doi: 10.18653/v1/2020.acl-demos.22 2

[30] S. Jentzsch and C. Turan. Gender bias in BERT - measuring and analysing biases through sentiment rating in a realistic downstream classification task. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pp. 184–199. Association for Computational Linguistics, Seattle, Washington, July 2022. doi: 10.18653/v1/2022.gebnlp-1.20 1

[31] A. Kasirzadeh and A. Smart. The use and misuse of counterfactuals in ethical machine learning. In *Proc. FAccT '21*, p. 228–236, 2021. doi: 10.1145/3442188.3445886 2, 7

[32] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, p. 1–14, 2020. doi: 10.1145/3313831.3376219 2, 3, 4

[33] D. Kaushik, E. Hovy, and Z. C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data, 2019. doi: 10.48550/ARXIV.1909.12434 2

[34] M. T. Keane, E. M. Kenny, E. Delaney, and B. Smyth. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques. *ArXiv*, abs/2103.01035, 2021. 2

[35] U. Kuhl, A. Artelt, and B. Hammer. Keep your friends close and your counterfactuals closer: Improved learning from closest rather than plausible counterfactual explanations in an abstract setting. In *Proc. FAccT '22*, p. 2125–2137, 2022. doi: 10.1145/3531146.3534630 2

[36] J. Lalor, Y. Yang, K. Smith, N. Forsgren, and A. Abbasi. Benchmarking intersectional biases in nlp. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3598–3609, 2022. 1

[37] J. Li, X. Chen, E. Hovy, and D. Jurafsky. Visualizing and understanding neural models in nlp, 2015. doi: 10.48550/ARXIV.1506.01066 6

[38] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He. A survey on text classification: From traditional to deep learning. *ACM Trans. Intell. Syst. Technol.*, 13(2), apr 2022. doi: 10.1145/3495162 1

[39] Q. V. Liao, M. Pribić, J. Han, S. Miller, and D. Sow. Question-driven design process for explainable ai user experiences, 2021. doi: 10.48550/ARXIV.2104.03483 2

[40] Q. V. Liao and K. R. Varshney. Human-centered explainable ai (xai): From algorithms to user experiences, 2021. doi: 10.48550/ARXIV.2110.10790 2, 4

[41] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *ArXiv*, abs/1705.07874, 2017. 1, 2

[42] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. doi: 10.1016/j.artint.2018.07.007 4, 7

[43] T. Miller, P. Howe, and L. Sonenberg. Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. 2017. doi: 10.48550/ARXIV.1712.00547 2

[44] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao. Deep learning–based text classification: A comprehensive review. *ACM Comput. Surv.*, 54(3), apr 2021. doi: 10.1145/3439726 1

[45] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In *Proc. FAT* '19*, p. 220–229, 2019. doi: 10.1145/3287560.3287596 1

[46] R. K. Mothilal, A. Sharma, and C. Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proc. FAT* '20*, p. 607–617, 2020. doi: 10.1145/3351095.3372850 1, 2

[47] C. M. Myers, E. Freed, L. F. L. Pardo, A. Furqan, S. Risi, and J. Zhu. Revealing neural network bias to non-experts through interactive counterfactual examples. *ArXiv*, abs/2001.02271, 2020. 2

[48] M. Nadeem, A. Bethke, and S. Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5356–5371, 2021. doi: 10.18653/v1/2021.acl-long.416 1

[49] J. Nee, G. M. Smith, A. Sheares, and I. Rustagi. Linguistic justice as a framework for designing, developing, and managing natural language processing tools. *Big Data & Society*, 9(1):20539517221090930, 2022. doi: 10.1177/20539517221090930 1, 7

[50] M. Nourani, S. Kabir, S. Mohseni, and E. D. Ragan. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. *Proc. AAAI HCOMP*, 7(1):97–105, Oct. 2019. 4

[51] H. Orgad, S. Goldfarb-Tarrant, and Y. Belinkov. How gender debiasing affects internal model representations, and why it matters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2602–2628, 2022. 1

[52] B. Paranjape, M. Lamm, and I. Tenney. Retrieval-guided counterfactual generation for qa, 2021. doi: 10.48550/ARXIV.2110.07596 2

[53] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019. 2

[54] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proc. KDD '16*, p. 1135–1144, 2016. doi: 10.1145/2939672.2939778 1, 2

[55] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proc. ACL '20*, pp. 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442 2

[56] K. Sakaguchi, R. Le Bras, C. Bhagavatula, and Y. Choi. Winogrande: An adversarial winograd schema challenge at scale. *Proc. AAAI'20*, 34(05):8732–8740, Apr. 2020. doi: 10.1609/aaai.v34i05.6399 2

[57] A. M. Salih, Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva, S. E. Petersen, K. Lekadir, and G. Menegaz. A perspective on explainable artificial intelligence methods: Shap and lime. *Advanced Intelligent Systems*. doi: 10.1002/aisy.202400304 2

[58] R. Shang, K. J. K. Feng, and C. Shah. Why am i not seeing it? understanding users' needs for counterfactual explanations in everyday recommendations. In *Proc. FAccT '22*, p. 1330–1340, 2022. doi: 10.1145/3531146.3533189 2

[59] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings IEEE Symposium on Visual Languages*, pp. 336–343, 1996. doi: 10.1109/VL.1996.545307 6

[60] A. Silva, P. Tambwekar, and M. Gombolay. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proc. NAACL '21*, pp. 2383–2389, 2021. doi: 10.18653/v1/2021.naacl-main.189 1, 7

[61] D. Slack, S. Hilgard, H. Lakkaraju, and S. Singh. Counterfactual explanations can be manipulated. In *NeurIPS*, 2021. 2

[62] T. Spinner, U. Schlegel, H. Schäfer, and M. El-Assady. explainer: A visual analytics framework for interactive and explainable machine learning. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1064–1074, 2020. doi: 10.1109/TVCG.2019.2934629 1, 2

[63] I. Stepin, J. M. Alonso, A. Catalá, and M. Pereira-Fariña. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001, 2021. 2

[64] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. 2017. doi: 10.48550/ARXIV.1703.01365 1

[65] H. Suresh, S. R. Gomez, K. K. Nam, and A. Satyanarayan. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, 2021. doi: 10.1145/3411764.3445088 2, 4

[66] H. Suresh and J. Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, 2021. doi: 10.1145/3465416.3483305 1

[67] I. Tenney, J. Wexler, J. Bastings, T. Bolukbasi, A. Coenen, S. Gehrmann, E. Jiang, M. Pushkarna, C. Radebaugh, E. Reif, and A. Yuan. The language interpretability tool: Extensible, interactive visualizations and analysis for nlp models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 107–118, 2020. doi: 10.18653/v1/2020.emnlp-demos.15 1

[68] H. Vasconcelos, M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein, and R. Krishna. Explanations can reduce overreliance on ai systems during decision-making. *Proc. CSCW'23*, 7, apr 2023. doi: 10.1145/3579605 1, 7

[69] J. Vig. A multiscale visualization of attention in the transformer model, 2019. doi: 10.48550/ARXIV.1906.05714 2

[70] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard journal of law & technology*, 31:841–887, 04 2018. doi: 10.2139/ssrn.3063289 1, 2

[71] E. Wallace, J. Tuyls, J. Wang, S. Subramanian, M. Gardner, and S. Singh. Allennlp interpret: A framework for explaining predictions of nlp models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pp. 7–12, 2019. doi: 10.18653/v1/D19-3002 2

[72] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems, 2019. doi: 10.48550/ARXIV.1905.00537 1

[73] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim. Designing theory-driven user-centric explainable ai. In *Proc. CHI '19*, p. 1–15, 2019. doi: 10.1145/3290605.3300831 2

[74] Z. J. Wang, J. Wortman Vaughan, R. Caruana, and D. H. Chau. Gam coach: Towards interactive and user-centered algorithmic recourse. In *Proc. CHI'23*, 2023. doi: 10.1145/3544548.3580816 2

[75] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, C. Biles, S. Brown, Z. Kenton, W. Hawkins, T. Stepleton, A. Birhane, L. A. Hendricks, L. Rimell, W. Isaac, J. Haas, S. Legassick, G. Irving, and I. Gabriel. Taxonomy of risks posed by language models. In *Proc. FAccT '22*, p. 214–229, 2022. doi: 10.1145/3531146.3533088 1, 7

[76] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):56–65, 2020. doi: 10.1109/TVCG.2019.2934619 1, 2

[77] T. Wu, M. T. Ribeiro, J. Heer, and D. Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6707–6723, 2021. doi: 10.18653/v1/2021.acl-long.523 2

[78] Y. Yacoby, B. Green, C. L. Griffin Jr., and F. Doshi-Velez. "if it didn't happen, why would i change my decision?": How judges

respond to counterfactual explanations for the public safety assessment. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 10(1):219–230, Oct. 2022. doi: 10.1609/hcomp.v10i1.22001 1

[79] W. Yang, H. Le, S. Savarese, and S. C. H. Hoi. Omnixai: A library for explainable ai. *ArXiv*, 2022. doi: 10.48550/ARXIV.2206.01612 1, 2

[80] J. Yuan and E. Bertini. Context sight: model understanding and debugging via interpretable context. In *Proc. HILDA'22*, 2022. doi: 10.1145/3546930.3547502 2

[81] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 15–20. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-2003 1