Do LLMs Think Fast and Slow? A Causal Study on Sentiment Analysis

Zhiheng Lyu[*]
University of Hong Kong
zhlyu@cs.hku.hk

Zhijing Jin[*]
University of Toronto
zjin@cs.toronto.edu

Fernando Gonzalez
ETH Zürich
fer.adauto@gmail.com

Rada Mihalcea
University of Michigan
mihalcea@umich.edu

Bernhard Schölkopf
MPI for Intelligent Systems
bs@tue.mpg.de

Mrinmaya Sachan
ETH Zürich
msachan@ethz.ch

Abstract

Sentiment analysis (SA) aims to identify the sentiment expressed in a text, such as a product review. Given a review and the sentiment associated with it, this work formulates SA as a combination of two tasks: (1) a causal discovery task that distinguishes whether a review "primes" the sentiment (Causal Hypothesis C1), or the sentiment "primes" the review (Causal Hypothesis C2); and (2) the traditional prediction task to model the sentiment using the review as input. Using the peak-end rule in psychology, we classify a sample as C1 if its overall sentiment score approximates an average of all the sentence-level sentiments in the review, and C2 if the overall sentiment score approximates an average of the peak and end sentiments. For the prediction task, we use the discovered causal mechanisms behind the samples to improve LLM performance by proposing *causal prompts* that give the models an inductive bias of the underlying causal graph, leading to substantial improvements by up to 32.13 F1 points on zero-shot five-class SA.[1]

1 Introduction

Sentiment analysis (SA) is the task of identifying the sentiment y given a piece of text x. The field has a rich history originating from subjectivity analysis (Wiebe, 1994; Hatzivassiloglou and Wiebe, 2000), and developed rapidly with the availability of large opinionated online data such as reviews with ratings (Turney, 2002; Nasukawa and Yi, 2003; Zhang et al., 2015; Keung et al., 2020, *inter alia*).

Despite recent advances in large language models (LLMs), it is still challenging to address the fine-grained five-class SA (which corresponds to the five star ratings in most datasets) for document-level classification (Choi et al., 2020; Fei et al., 2023; Truică et al., 2021), due to the subtle nature of the task including aspects such as inter-aspect relations, commonsense reasoning, among others (Poria et al., 2023; Venkit et al., 2023).

In this paper, we propose a causally-informed solution for the SA task. Different from the approach of naïvely applying up-to-date LLMs, we leverage insights from causal inference to propose a reformulation for SA into two tasks, as in Figure 1: (1) a causal discovery task to identify the cause-effect relation between the review X and the sentiment Y, and (2) the traditional prediction task $f : x \mapsto y$ to model the sentiment using the review as input.

We first look into the causal discovery task. In the study of affect science (Salovey and Mayer, 2004; Barrett, 2006; Feinstein, 2013), language can be the cause of emotion (Satpute et al., 2013; Kassam and Mendes, 2013) – namely a review priming the following sentiment, i.e., the Causal Hypothesis C1 of $X \rightarrow Y$); or emotion can affect the use of language (Barrett, 2006) – namely sentiment priming the review as an ad-hoc justification for the emotion, i.e., the Causal Hypothesis C2 of $Y \rightarrow X$. These two processes might arise from the data annotation process (Jin et al., 2021), but is hard to discover post-hoc in existing datasets.

Given the possibility of both causal directions $X \rightarrow Y$ or $Y \rightarrow X$ in the SA data, we identify the actual underlying mechanism based on insights from psychology (Kahneman, 2011; Epstein, 1994). Specifically, we identify the correspondence of the above two causal mechanisms with the *Fast* and *Slow Thinking* systems (Kahneman, 2011): (1) a review-driven sentiment (as in C1) largely resembles the Slow Thinking process applying reasoning based on evidence, and (2) the process of first coming up with the sentiment and then justifying it by a review (as in C2) conforms to Fast Thinking. Given this correspondence, we apply the peak-end rule

---

Task 1. Causal Discovery

Two Possible Causal Directions:

C1 Review causes Sentiment
$X \rightarrow Y$ — Corresponds to — **Slow Thinking**
Overall = Average of the Emotion Arc

C2 Sentiment causes Review
$Y \rightarrow X$ — Corresponds to — **Fast Thinking**
Overall = Peak-End of the Emotion Arc

Psychology Theories (§3)

Applying the *Peak-End Rule* to discover the causal direction

Review
- This was a great spot to take a break from it all and just people watch.
- We sat at the bar facing the casino and we were entertained the whole time.
- The mini grilled cheese (appetizer) was fantastic.
- It came with a tomato based dipping sauce that was the perfect compliment to the bite sized wedges.
- Tip - ask for two dipping sauces because one just won't do.

Emotion Arc

Peak-End Rule (from *Thinking Fast and Slow*):
- For slow thinking, overall sentiment ≈ **average** feeling
- For fast thinking, overall sentiment ≈ average of the **peak** (most intensive) and **end** feelings

Label: 50% (which is closer to Mean than to Peak-End)
Mean: 46% (by slow thinking)
Peak-End: 13% (by fast thinking)

Task 2. Improving LLM Performance on SA

1. LLMs perform better on C2 data (§4.1)
2. We can give inductive bias to LLMs by *causal prompts* (§4.2)
3. LLMs follow some mechanisms with causal prompts, but not perfectly (§4.3)

Causal Prompts (§4.2)

C1 I **initially composed the following feedback**: {review} After carefully considering the facts, my final rating was ___

C2 I **initially selected a star rating**, and then provided the following explanations in my review: {review} The review clarifies why I gave a rating of ___
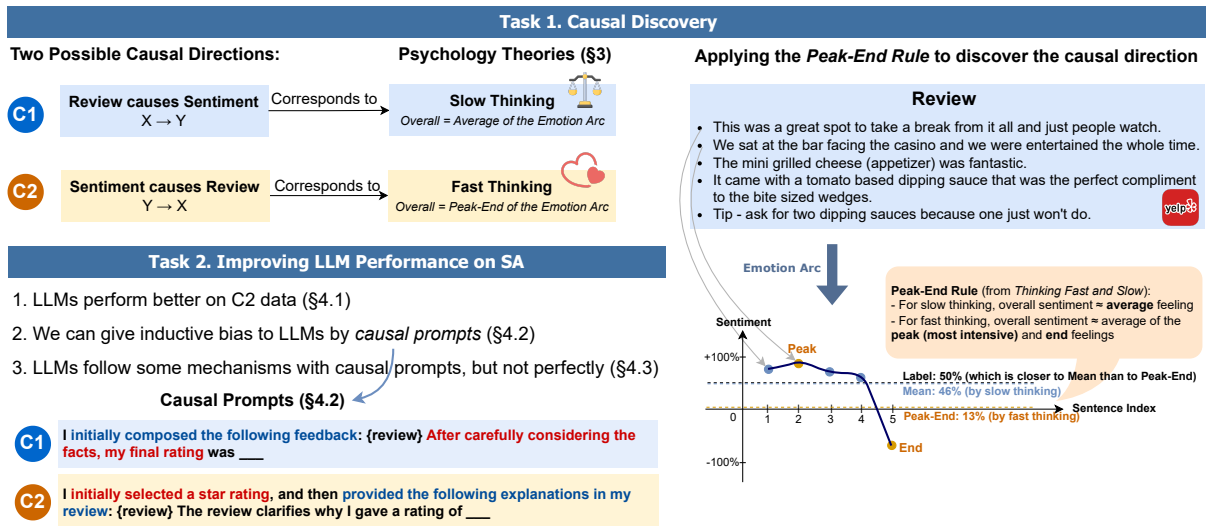
Figure 1: An overview of the paper structure, where we first investigate the causal discovery task, and then use it to improve LLM performance. For each document-level text review, we parse its *emotion arc* consisting of the sentiment of each sentence in the review, and then use the peak-end rule (Kahneman et al., 1993; Kahneman, 2011) to identify whether the overall sentiment is an average of the arc (corresponding to *Slow Thinking*), or an average of the peak and end sentiments (corresponding to *Fast Thinking*).

from psychology (Kahneman et al., 1993; Kahneman, 2011). As shown in the right part of Figure 1, we classify a sample as C1 if its overall sentiment score approximates an average of all the sentence-level sentiments in the review, and as C2 if the overall sentiment score approximates an average of the peak and end sentiments.

Based on the identified causal mechanism behind SA data from the causal discovery task, we further explore how it can improve prediction performance in the era of LLMs. Existing literature highlights "causal alignment," namely to align the prediction direction along the underlying causal direction (Jin et al., 2021; Schölkopf, 2022; Schölkopf et al., 2021), but to our knowledge we are the first to explore how causal alignment improves model performance of SA in the era of LLMs. Specifically, we answer three subquestions: (Q1) If using the standard SA prompt, do models perform differently on C1/C2 data? (Q2) Does it help if we make the prompt aware of the underlying causality, i.e., use *causal prompts*? And (Q3) When prompted causally, do LLMs mechanistically understand the corresponding causal processes?

Our empirical results show that under the standard prompt, LLMs perform better on data corresponding to the C2 causal process. Moreover, causal prompts aligned with the causal direction of the data can substantially improve the performance of zero-shot SA. Finally, we apply mechanistic inter-

pretability methods to probe the models, and find that there is still improvement space for LLMs to correctly grasp the essence of the two causal processes. In summary, the contributions of this paper are as follows:

1. We propose the dual nature of SA as a combination of two tasks: a causal discovery task, and a prediction task.
2. For causal discovery, we ground the two possible causal processes in psychology, and use the peak-end rule to identify them.
3. For the prediction task, we inspect existing LLMs' performance on data corresponding to the two underlying causal processes, and design *causal prompts* to improve model performance by up to 32.13 F1 points.

2 Problem Formulation of SA

In this section, we formulate SA as a combination of two tasks: the traditional prediction task in NLP and the causal discovery task in statistics, which we will introduce in the following.

2.1 The Prediction Task (in NLP)

SA is a prediction task to identify the sentiment y given a piece of text x. We adopt the setup in most existing SA datasets (Maas et al., 2011; Zhang et al., 2015; Keung et al., 2020), where the text x is a review consisting of n sentences $(t_1, \ldots, t_n)$, and the label y is a sentiment score corresponding

to the star rating of the review in 1 (most negative), 2, ..., 5 (most positive).

2.2 The Cause-Effect Discovery Task (in Statistics)

As a separate problem, there is an established task in causal discovery, the causal-effect problem (see the review by Janzing, 2019), which aims to tell the cause from effect using only observational data. Its formal formulation is as follows: Suppose we have an i.i.d. dataset $\mathcal{D} := \{(x_i, y_i)\}_{i=1}^n$ containing n observational data pairs of the two variables, X and Y. The task is to *infer whether X causes Y (i.e., $X \to Y$), or Y causes X (i.e., $Y \to X$)*, if one out of the two is true. In causality, "$\to$" indicates the directional causal relation between two variables. The two hypotheses can also be expressed in their equivalent structural causal models (SCMs; Pearl et al., 2000) as introduced in Peters et al. (2017):

$$\text{Causal Hypothesis 1 (C1): } X \to Y \qquad (1)$$
$$\Leftrightarrow Y := f_Y(X, N_Y) \text{ with } N_Y \perp X , \quad (2)$$

$$\text{Causal Hypothesis 2 (C2): } Y \to X \qquad (3)$$
$$\Leftrightarrow X := f_X(Y, N_X) \text{ with } N_X \perp Y , \quad (4)$$

where N_i is an unobserved noise term orthogonal to the input distribution.

2.3 Causality and NLP Model Performance

For many years, causality and machine learning have been two separate domains on their own. Recently, researchers started to think about how the causal knowledge of the data can improve machine learning performance on the prediction task, especially for the two variable cause-effect case (Schölkopf et al., 2012; Jin et al., 2021; Ni et al., 2022). The essence of this line of research is that causality makes the two learning tasks $x \mapsto y$ and $y \mapsto x$ asymmetric, as one function's prediction direction *aligns with* the ground-truth causal direction behind the two random variables, and another contradicts. We call this phenomenon "*causal alignment,*" or "direction match," of the prediction task and the causality.

To contrast the contribution of our work, we review the previous literature on causal alignment, which only shows its effect on the performance of trained-from-scratch machine learning models, without any indications in the era of LLMs:

1. Causal alignment makes a model more robust against **covariant shifts** (Jin et al., 2021; Schölkopf, 2022; Schott et al., 2018, *inter alia*)

2. Semi-supervised learning (SSL) only works under causal misalignment, as the cause variable contains no information about the mechanism, but the effect variable does. So in the misaligned case, additional P_X (i.e., the effect variable) helps **SSL** (Schölkopf et al., 2012; Jin et al., 2021).

3. Learning a causally-aligned model induces **less Kolmogorov complexity** (a more minimal description length) than the causally-misaligned model on the same X-Y data (Jin et al., 2021; Janzing and Schölkopf, 2010)

4. Causal alignment significantly affects model performance in **supervised learning**, in the case of machine translation (Ni et al., 2022).

All the above findings are drawn under the training condition that we can isolate the training data to be only of one causal direction. In the era of LLMs, we have seen substantial differences: (1) the training data can be a mixture of both causal directions, (2) the operationalization of the prediction task is through prompting, but no longer a separate model for each direction, and, (3) in general, research has shifted to designing better prompts for already pre-trained models in their inference mode.

Given these changes, we use the rest of the paper to address the following research questions:

1. What is the causal direction in SA? (Section 3)
2. Can causal alignment help us improve SA prompts in the era of LLMs? (Section 4)

3 Causal Discovery: Does Sentiment Cause the Review, or Vice Versa?

3.1 Problem Setup

As mentioned previously, the setup of the bivariate causal discovery problem is to infer whether X causes Y (C1), or Y causes X (C2), based on a dataset $\mathcal{D} := \{(x_i, y_i)\}_{i=1}^n$ containing only observational data of the joint distribution.

Challenges The common paradigm to check causal discovery results is to generate simulated data, of which the ground truth causal graph is known (Zhang and Hyvärinen, 2009; Spirtes and Zhang, 2016). However, in the context of the es-

tablished SA datasets, such as Yelp (Zhang et al., 2015), Amazon (Keung et al., 2020), and App Review (Grano et al., 2017), we would not be able to track each individual user and survey their original causal process when composing the review and the rating. Another solution would also be difficult, as it would require SA to abandon all the above well-established datasets, and meticulously collect new data while surveying the users' underlying causal process.

Our Approach In the context of our work, we propose that there are still rich findings that we could derive from the observation-only data in the existing datasets, without interviewing or conducting new costly data collection.

The key to our approach is the psychology theories of the two causal processes, as the relation between sentiment and text has been well-studied and verified by randomized control trials (RCTs), among many other experiments. In the rest of the section, we first introduce in Section 3.2 the psychology theories of fast and slow thinking, followed by the Peak-End Rule as the quantitative signal. Then, we operationalize the theory with computational techniques in Section 3.3, and the report findings on three different SA datasets in Section 3.4.

3.2 Psychological Processes Underlying Sentiment Processing

Two Systems of Emotional Responses In psychology, the bifurcation into System 1 and System 2 in human decision-making, including sentiment processing, has garnered substantial empirical support (Kahneman, 2011; Epstein, 1994).

System 1, or the "*Fast Thinking*" system, operates involuntarily, effortlessly, and without conscious awareness. It is often optimized in evolution to provide rapid responses to environmental stimuli (LeDoux, 1998), and guides most of our daily cognitive processing (Kahneman, 2011), and emotional responses such as fear or joy (Zajonc, 1980).

Conversely, System 2, often termed as "*Slow Thinking*," is deliberate, slower, and more rational, requires more conscious effort (Kahneman, 2011), and allows for self-regulation and thoughtful consideration before making decisions (Baumeister et al., 1998). The interplay between these systems influences everything from mundane to critical decisions, highlighting the complexity of human emotional and cognitive processing (Kahneman and

Frederick, 2002; Kahneman, 2011).

Correspondence to the Two Causal Processes There is a nice correspondence between the fast/slow thinking systems and our two causal hypotheses. As mentioned previously, the Causal Hypothesis 2 (C2) posits $Y \rightarrow X$, where the sentiment Y causes the review X, which aligns well with the *Fast Thinking* system (Kahneman, 2011; LeDoux, 1998), as it rapidly generates an emotional reaction Y, and then writes text to justify it Y. On the other hand, the Causal Hypothesis 1 (C1) refers to the case where $X \rightarrow Y$, namely the review X causing the sentiment Y. It is an instance of *Slow Thinking* (Baumeister et al., 1998; Kahneman and Frederick, 2002), which deliberately uses conscious efforts to list out the up- and downsides of an experience in the review X, and come up with a thoughtful final decision as the rating Y.

Quantitative Signals of the Two Processes In sentiment processing, an evidence for the two processes is the famous Kahneman et al. (1993) study illustrating the *Peak-End Rule* of how individuals recall and evaluate past emotional experiences, which we show in Figure 1. As we know, fast thinking is prone to systematic biases and errors in the judgment (Tversky and Kahneman, 1974), and the Kahneman et al. (1993) study provides important quantitative results showing that, in the Fast Thinking system, people's emotional memories of an experience are disproportionately influenced by its most intense point (the "peak") and its conclusion (the "end"), rather than by the average experience as in the Slow Thinking system. The important role of peak and end for the fast thinking system implies that it is the intensity of specific moments that dominate memory and judgment.

3.3 Operationalizing the Theory

We summarize the previous psychological insights in the upper left part of Figure 1, where the Causal Hypothesis 1 corresponds to taking the average of all emotional experiences mentioned in the review X for the sentiment Y, and the Causal Hypothesis 2 uses the peak and end emotions in the review X to derive the sentiment Y. In this section, we introduce a formalization of the theory, and suggest signals to distinguish the two causal hypotheses.

Emotion Arc To capture the aforementioned trajectory of emotional experiences, we use the concept of the *emotion arc* (Reagan et al., 2016), an

	Yelp			Amazon			App Review		
	All	C1	C2	All	C1	C2	All	C1	C2
# Samples	34,851	19,557 (56%)	15,294 (44%)	2,582	1,393 (54%)	1,189 (46%)	9,696	3,809 (39%)	5,887 (61%)
# Sents/Review	11.11	11.30	10.87	6.70	6.62	6.80	6.34	6.33	6.35
# Words/Sent	15.53	15.55	15.49	11.04	11.28	10.77	10.53	10.90	10.29
Vocab Size	64,864	48,889	44,826	10,271	7,609	7,049	20,248	12,773	15,400
Avg Sentiment	2.93	2.74	3.18	2.94	2.82	3.07	2.9	2.72	3.02
Avg λ_1	3.78	2.97	4.83	3.77	3.10	4.55	6.03	5.18	6.58
Avg λ_2	4.48	6.05	2.48	4.21	5.72	2.45	5.10	7.96	3.26

Table 1: Statistics of the entire datasets and their C1 and C2 subsets for Yelp, Amazon, and App Review. We can see that a roughly balanced number of reviews aligning with the C1 and C2 processes.

example of which we visualize in Figure 1. Contextualizing it in the task of SA, we formally define an emotion arc of the review as follows. Given a review x consisting of n sentences $(t_1, \ldots, t_n)$, we identify the sentiment for each of them, thus obtaining a series of sentiment labels $(s_1, \ldots, s_n)$. We denote this series as the emotion arc $\boldsymbol{e} := (s_1, \ldots, s_n)$ of the review.

The Two Causal Processes Provided the notion of the emotion arc $\boldsymbol{e} := (s_1, \ldots, s_n)$ for a review x, we formulate the sentiment labels corresponding to the two causal processes as follows:

Slow Thinking (Causal Process 1):

$$\hat{y}_{\text{avg}} = \frac{1}{n} (s_1 + \cdots + s_n) \ , \tag{5}$$

$$\lambda_1 = |y - \hat{y}_{\text{avg}}| \ , \tag{6}$$

Fast Thinking (Causal Process 2):

$$\hat{y}_{\text{peakEnd}} = \frac{1}{2} (\text{Peak}(s_1, \ldots, s_n) + s_n) \ , \tag{7}$$

$$\lambda_2 = |y - \hat{y}_{\text{peakEnd}}| \ , \tag{8}$$

where λ_i indicates the alignment of the actual sentiment y with the Causal Process i, and $\text{Peak}(\cdot)$ selects the sentiment with the strongest intensity by its distance from the neutral sentiment 3, which is the middle point among the sentiment range 1–5, i.e., $\text{Peak}(s_1, \ldots, s_n) := s_{\text{argmax}_i |s_i - 3|}$.

Here, we interpret λ_i as an indicator for each causal process, where a small value (with the best value being zero) implies the alignment with the process i. We show two examples in Table 2, one aligning well with the Causal Process C1 with a small λ_1, and another aligning well with the Causal Process C2 with a small λ_2.

3.4 Findings on SA Datasets

Dataset Setup We adopt three commonly used datasets in SA: Yelp (Zhang et al., 2015), Amazon (Keung et al., 2020), and App Review (Grano

---

Example of a C1-Dominant Review
· This was a great spot to take a break from it all and just people watch. $s_1 = 4.57$
· We sat at the bar facing the casino and we were entertained the whole time. $s_2 = 4.67$
· The mini grilled cheese (appetizer) was fantastic. $s_3 = 4.53$
· It came with a tomato based dipping sauce that was the perfect compliment to the bite sized wedges. $s_4 = 4.20$
· Tip - ask for two dipping sauces because one just won't do. $s_5 = 1.60$
Stars y: 4
Psychology Scores ($\downarrow$): $\lambda_1 = 0.0884 < \lambda_2 = 0.8683$

Example of a C2-Dominant Review
· I read the reviews and should have steered away... but it looked interesting. $s_1 = 3.72$
· Salad was wilted, menus are on the wall, with no explanation so you are ordering blind, service was NOT with a smile from the bartender to the waitress, to the server who helped the waitress, and the waitress never checked back to see how everything is. $s_2 = 2.20$
· Terribly overpriced for what you get, and as an Italian, this does not even pass for a facsimile thereof! $s_3 = 1.45$
· Stay away for sure. $s_4 = 1.85$
· I only gave them one star, as I had to fill something in, they should get no stars! $s_5 = 1.32$
Stars y: 1
Psychology Scores ($\downarrow$): $\lambda_1 = 1.1827 > \lambda_2 = 0.3647$

Table 2: Examples of C1- and C2-dominant reviews.

---

et al., 2017). For the Amazon data, we concatenate each review's title with its text. Since the model performance on many binary classification datasets is saturated (Poria et al., 2023; Yang et al., 2019), we use the 5-way classification version of the SA datasets when applicable.

Since we need to utilize the emotion arc, we keep only reviews with at least five sentences, after sentence tokenization using the Spacy package (Honnibal and Montani, 2017). We apply this filtering above on the test set of the Yelp dataset, the English test of Amazon, and the unsplit entire dataset of App Review. We report the statistics of remaining samples in Table 1.

To obtain the emotion arcs, we calculate the sentiment score of each sentence produced by the

		Random	GPT-2 XL	LLaMa-7B	Alpaca-7B	GPT-3	GPT-3.5	GPT-4
F1	Overall	19.82 ±2.07	10.23 ±4.12	31.78 ±5.32	46.01 ±5.35	52.71 ±1.73	57.98 ±5.11	59.54 ±4.69
	C1 Subset	21.36 ±2.26	5.80 ±3.11	27.30 ±4.73	37.77 ±7.66	43.96 ±2.93	**58.64** ±1.48	58.62 ±2.54
	C2 Subset	20.43 ±2.95	**16.37** ±5.33	37.66 ±7.86	**55.82** ±4.02	**65.40** ±1.37	59.09 ±9.13	62.57 ±6.85
Accuracy	Overall	19.78 ±2.07	23.06 ±2.10	39.28 ±5.07	47.72 ±4.19	53.22 ±1.35	58.36 ±4.13	59.84 ±4.17
	C1 Subset	20.61 ±2.23	16.18 ±1.59	36.55 ±4.05	42.14 ±5.26	43.61 ±2.89	**59.89** ±1.09	59.62 ±1.96
	C2 Subset	18.86 ±2.78	**30.79** ±2.68	42.33 ±7.24	53.93 ±3.91	**63.93** ±1.28	56.66 ±8.08	60.08 ±7.05

Table 3: Performance of different models on the five-class classification of Yelp-5. We use five paraphrases for the prompt (in Appendix A.4), and report the average performance with the standard deviation.

		Random	GPT-2 XL	LLaMa-7B	Alpaca-7B	GPT-3	GPT-3.5	GPT-4
F1	Data=C1, Prompt=C1	20.47 ±2.47	6.12 ±2.77	55.16 ±7.16	52.74 ±5.04	38.36 ±6.66	60.62 ±3.24	54.23 ±4.17
	Data=C1, Prompt=C2	20.26 ±2.31	15.98 ±3.59	36.22 ±8.95	35.10 ±5.40	54.44 ±1.64	52.85 ±6.01	58.58 ±3.33
	Data=C2, Prompt=C1	22.35 ±3.02	31.98 ±8.66	56.69 ±8.46	54.74 ±14.26	74.64 ±3.06	**78.18** ±1.21	72.52 ±3.68
	Data=C2, Prompt=C2	20.35 ±2.18	**48.50** ±7.66	**66.82** ±7.90	**71.22** ±3.99	77.09 ±1.33	78.16 ±1.81	**76.80** ±1.36
Acc	Data=C1, Prompt=C1	19.60 ±2.67	12.96 ±2.91	58.23 ±5.30	55.81 ±4.00	43.16 ±6.20	60.39 ±3.25	54.06 ±3.97
	Data=C1, Prompt=C2	19.68 ±2.46	22.61 ±5.73	43.07 ±8.18	41.45 ±4.06	56.36 ±1.94	53.11 ±5.89	58.68 ±3.45
	Data=C2, Prompt=C1	20.97 ±3.19	43.51 ±6.30	57.54 ±7.21	54.82 ±12.59	76.60 ±3.23	77.38 ±1.43	70.69 ±3.93
	Data=C2, Prompt=C2	19.03 ±2.18	**51.59** ±5.09	**68.16** ±9.24	**71.83** ±4.33	**76.70** ±1.35	**78.92** ±1.31	75.38 ±1.70

Table 4: Performance on Yelp using the *causal prompts* on the two causal subsets. We report the average performance across the five paraphrases for each prompt, with the standard deviation.

`sentiment-analysis` pipeline[2] from Huggingface (Wolf et al., 2020).

Causal Discovery For each input sample, we process them as in Table 2, namely first obtaining the sentence-level sentiments to form the emotion arc, and then calculating the alignment scores λ_1 and λ_2 for each causal process, respectively. We consider an example as *dominated* by the causal process C_i if the alignment score λ_i is more optimal than the other. We report the resulting statistics in Table 1. For each dataset, we describe their overall statistics, as well as the statistics of data with the underlying causal process of C1, and that of C2. We can see that Yelp and Amazon have an almost balanced split of C1 and C2, while App Review has 61% C2 data compared to 39% C1 data. See Appendix B.2 for an additional visualization of the λ_1-λ_2 distribution across the 1K data points.

4 How to Improve Sentiment Classifiers with Causal Alignment?

Using our proposed causal discovery method, we have identified two distinct subsets with their corresponding causal processes C1 and C2. Now, we address the last practical question proposed in Section 2.3:

> *Can causal alignment help us improve SA in the era of LLMs?*

Specifically, we take the commonly used approach

in the era of LLMs, i.e., prompting pre-trained LLMs for the SA task, and look into how alignment with the underlying causal process could help SA performance. We answer the following three subquestions in this section:

Q1. Using the standard SA prompt, do models perform differently on C1/C2 data? (Section 4.1)

Q2. Does it help if we make the prompt aware of the underlying causality, i.e., use causal prompts? (Section 4.2)

Q3. When prompted causally, do LLMs really understand the causal processes? (Section 4.3)

4.1 Q1: Do Models Perform Differently on C1/C2 Data?

Experimental Setup The first question is whether models perform differently on data with the causal nature of C1 or C2. We use the subsets identified by our psychologically-grounded causal discovery, and test a variety of available autoregressive LLMs, including the open-weight GPT-2 (Radford et al., 2019), LLaMa (Touvron et al., 2023), and Alpaca (Taori et al., 2023); as well as the closed-weight models with OpenAI API, the instruction-tuned GPT-3 (text-davinci-002) (Brown et al., 2020; Ouyang et al., 2022), GPT-3.5 (gpt-3.5-turbo-0613), GPT-4 (gpt-4-0613) (OpenAI, 2023). We also add a random baseline which uniformly samples the label space for each input.

We use the standard prompt formulation for SA in the format of "`[Instruction] Review`

---

[2] https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english. See details in Appendix A.2.

`Text: {x}\n Label:`". The experiments are on a set of randomly selected 1K samples from the test set of Yelp-5 (Zhang et al., 2015), due to the time- and cost-expensive inference of the above LLMs. (E.g., LLaMa/Alpaca takes 96 GPU hours to run.) See more experimental details in Appendix A.1.

Results We show the performance of the six LLMs in Table 3, and report the F1 and accuracy across the five-class classification on Yelp-5. We can see that the existing LLMs perform the best on the subset with the causal process C2, implying that the decision pattern of LLMs is closer to the Fast Thinking system, which takes the peak-end average of the emotion arc.

4.2 Q2: Do Causal Prompts Help?

Designing *Causal Prompts* Inspired by the fact that models perform differently on C1/C2 data, our next question is, will it help if we directly give a hint to the LLMs about the underlying causal graph?

	Prompt Design
C1	As a customer writing a review, I initially *composed* the following feedback: "`[review]`"
	After carefully considering the facts, I selected a star rating from the options of "1", "2", "3", "4", or "5". My final rating was:
C2	As a customer writing a review, I initially *selected* a star rating from the options "1", "2", "3", "4", and "5", and then provided the following explanations in my review: "`[review]`"
	The review *clarifies* why I gave a rating of

Table 5: Causally-aware prompts describing the SA task in contexts with the C1 and C2 causal graphs.

To this end, we propose the idea of *causal prompts*, which are prompts that describe the causal story behind the input and output variables. We list our designed prompts for the C1 and C2 stories in Table 5.

Results We report the performance for all combinations of the dataset natures and prompt natures in Table 4, where we find that the most-performant setting uses Prompt C2 on the data subset with the same causal nature, C2. This alignment leads to the best performance across almost all models by both F1 and accuracy. On the C2 data, we also see that Prompt C2 outperforms the standard SA prompt in Table 3 by a substantial margin, such as 32.13 F1 points increase for GPT-2, and 14.23 F1 points increase for GPT-4.

However, although Prompt C2 shows a strong performance, the other causal prompt, i.e., Prompt C1, does not always help the data subset C1 in all cases, from which we raise a further question – how well do LLMs really mechanistically understand our prompts? We explore this question in the next section.

4.3 Q3: Can LLMs Correctly Capture the Causal Stories in the Prompts?

Although the proposal of the two causal prompts is intuitive for humans, we still need to inspect whether LLMs are able to understand them correctly.

Method Mechanistically, for a model to solve SA for the causal process C1 correctly, it needs to treat the sentence-level sentiments across all sentences *equally*; and for a model to solve SA for the causal process C2 correctly, it needs to pay *more* attention to the peak and end sentiments on the emotion arc.

Targeting the two mechanisms, we use causal tracing (Meng et al., 2022) to attribute the final sentiment prediction to the source sentences in the input. Briefly, causal tracing uses causal mediation analysis (Pearl, 2001) to quantify the causal contribution of the internal neuron activations of a model to its final prediction (Vig et al., 2020). We use causal tracing to inspect the causal effects of the hidden states on the model prediction, using the open-weight models, LLaMa and Alpaca. We use the causal effects of the first-layer neurons for each sentence, which we aggregate to obtain the final prediction. See implementation details in Appendix A.3.

Results We plot the causal attribution results of how much each sentence contributes to the final prediction in Figure 2. Here, the ideal behavior of the models is that Prompt C1 should trigger *uniform* attention over all the sentences, which is roughly observed through the more even shades of color of the "Prompt C1" row than the "Prompt C2" row in Figure 2 in the row of "Prompt C1".

On the other hand, Prompt C2 should trigger *more* attention to the sentences corresponding to the peak and end sentiments. For this, we see the models have high attention to the middle sentence, as in the "Prompt C2" row in Figure 2. The average causal effect of the peak sentence on predictions under Prompt C2 is 3% higher than the mean effect for the LLaMa model, and 39% higher for the
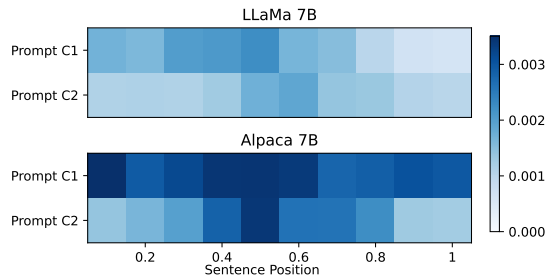
Figure 2: Causal attribution in LLaMa-7B and Alpaca-7B, showing how much each sentence contributes to the prediction probability.

Alpaca model. This aligns with our expectation that the peak sentence would have a high contribution. Nonetheless, note that no model sufficiently attends to the end sentence under Prompt C2. This implies that they do not fully grasp the expected contribution pattern of the peak-end rule, missing the significant role of the end sentence.

5 Related Work

SA The task of SA aims to identify the sentiment given a piece of text. It has a rich history originating from subjectivity analysis (Wiebe, 1994; Hatzivassiloglou and Wiebe, 2000), and developing rapidly with the availability of large opinionated online data such as reviews with star ratings (Turney, 2002; Nasukawa and Yi, 2003; Zhang et al., 2015; Keung et al., 2020, *inter alia*). Most literature on SA focuses on building computational models, from using traditional linguistic rules (Hatzivassiloglou and McKeown, 1997; Choi and Cardie, 2008), to the application of machine learning methods, from traditional naive bayes and support vector machines (Pang et al., 2002; Moraes et al., 2013; Tan et al., 2009), to early deep learning models (Socher et al., 2013; Kim, 2014; Xing et al., 2020), and finally entering the era of LLMs (Hoang et al., 2019; Raffel et al., 2020; Yang et al., 2019).

Psychology and Affective Science In the study of emotion, or affect science (Salovey and Mayer, 2004; Barrett, 2006; Feinstein, 2013), previous work finds that not only the emotion people perceive influences or prime how they communicate in the moment (Barrett, 2006), but language can also influence emotion, which can be observed in functional magnetic neuroimaging (Satpute et al., 2013), and also experiments showing the act of self-reporting the emotion in writing can change the physical reaction to the emotion (Kassam and Mendes, 2013). In his seminal work, Kahneman

(2011) uses the two systems of thinking to reveal the mechanisms of how people come up with their sentiment, where fast thinking conforms to the peak-end rule (Kahneman et al., 1993), and slow thinking is more reflective of the overall sentiment.

Cause-Effect Distinction Distinguishing the cause from effect based on observational data is a long-standing and fundamental problem in causality (Hoyer et al., 2008; Zhang and Hyvärinen, 2009; Janzing, 2019). Existing methods to address this problem are based on statistics (Hoyer et al., 2008; Peters et al., 2010; Shajarisales et al., 2015; Mooij et al., 2014), physics (Janzing, 2007; Janzing et al., 2016), information theory (Janzing et al., 2012; Chaves et al., 2014; Mejia et al., 2022), and algorithmic complexity (Janzing and Schölkopf, 2010; Jin et al., 2021). However, we are the first to look at the rich nature of NLP datasets, and directly approach the difference in the causal and anticausal mechanisms grounded in interdisciplinary insights.

As for our causal prompts, the most similar studies are the non-causally-grounded explorations for prompt tuning, such as by varying the patterns of masked language modeling (Schick and Schütze, 2022) and using the noisy channel method (Min et al., 2022). However, these studies are not aware of the underlying causal processes, thus neglecting the connection of prompts with the causal nature of data, and also the explicit causal story of the sentiment-review relation.

6 Conclusion

In conclusion, we have formulated the task of SA into a prediction problem and a causal discovery problem. We first identified the cause-effect relation among existing SA datasets, namely whether the review primes the rating, or the sentimental judgment primed the review writing process. To achieve this causal discovery, we obtain insights from existing psychology studies, namely aligning the above two causal processes with the famous Fast Thinking and Slow Thinking systems, with their distinct qualitative signals. Given the causal understanding of the dataset, we further improve the performance of LLMs on SA using our proposed causal prompts. Our research paves the way for more causally-aware future research in SA.

Limitations and Future Work

This study has several limitations. First, the rapid progression of LLMs makes it challenging to keep up with all newly proposed models and architectures. Since our work covers only a set of recent LLMs at the time of this study, we encourage future research to apply our methods to additional LLMs and other SA datasets.

Although our study is grounded in well-established psychological theories, there remains the possibility that new theories could emerge, necessitating updates to the calculation of the λ values for the two causal processes. However, the causal processes identified in this work appear plausible, as evidenced by the effectiveness of the causally aligned prompts in improving language model performance.

Regarding the causal graph formulation, we focus on basic bivariate causal graphs, but future work could include more variables, such as confounders, mediators, and colliders.

The nature of this work is to introduce a paradigm shift for SA, and formulate the task differently. Therefore, we see lots of space for future extensions, such as to explore the causal nature of SA in different settings, different languages, and also aspect-based sentiment analysis (Pontiki et al., 2014; Xing et al., 2020; Hua et al., 2023).

Ethical Considerations

Regarding data concerns and user privacy, our study employs several established NLP datasets, and the examples we cite do not include sensitive user information.

Concerning potential stakeholders and misuse, this research primarily introduces a new perspective on the SA task. A possible negative impact concerns the general application of SA, which could be used to analyze user mentality for surveillance or fraudulent purposes. We acknowledge that studies on SA inherently involve these risks, and we firmly oppose the misuse of SA models in such contexts.

Acknowledgment

Author Contributions

For the idea formulation, Zhijing Jin drew inspirations from her previous work (Jin et al., 2021) and proposed to extend it to LLMs. During the internship of Zhiheng Lyu at ETH Zürich hosted under Prof Mrinmaya Sachan and mentored by Zhijing, Zhiheng first explored whether LLMs show a distinct fingerprint for causal and anticausal prompts in a workshop paper (Lyu et al., 2022). Then Zhijing got inspired by chats with psychology researchers in affect science that the phenomena of sentiment-primed text and text-primed sentiments can be grounded in actual psychology studies. Hence, we started to explore this paper's idea together.

Zhiheng Lyu conducted all the experiments and analyses, proposed various novel experimental designs, structured the storyline, and did tremendous hard work to make the paper happen.

Zhijing Jin closely mentored the project, proposed the initial idea as well as the Fast and Slow Thinking formulation, and wrote the entire current version of the paper.

References

Lisa Feldman Barrett. 2006. Solving the emotion paradox: Categorization and the experience of emotion. *Personality and social psychology review*, 10(1):20–46. 1, 8

RF Baumeister, E Bratslavsky, M Muraven, and DM Tice. 1998. Ego depletion: is the active self a limited resource? *Journal of Personality and Social Psychology*, 74(5):1252–1265. 4

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. 6

Rafael Chaves, Lukas Luft, Thiago O. Maciel, David Gross, Dominik Janzing, and Bernhard Schölkopf. 2014. Inferring latent structures via information inequalities. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence, UAI 2014, Quebec City, Quebec, Canada, July 23-27, 2014*, pages 112–121. AUAI Press. 8

Gihyeon Choi, Shinhyeok Oh, and Harksoo Kim. 2020. Improving document-level sentiment classification using importance of sentences. *Entropy*, 22(12):1336. 1

Yejin Choi and Claire Cardie. 2008. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 793–801. ACL. 8

Seymour Epstein. 1994. Integration of the cognitive and the psychodynamic unconscious. *American psychologist*, 49(8):709. 1, 4

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1171–1182, Toronto, Canada. Association for Computational Linguistics. 1

Justin S Feinstein. 2013. Lesion studies of human emotion and feeling. *Current opinion in neurobiology*, 23(3):304–309. 1, 8

Giovanni Grano, Andrea Di Sorbo, Francesco Mercaldo, Corrado A Visaggio, Gerardo Canfora, and Sebastiano Panichella. 2017. Software applications user reviews. 4, 5

Vasileios Hatzivassiloglou and Kathleen R McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*, pages 174–181. Association for Computational Linguistics. 8

Vasileios Hatzivassiloglou and Janyce M Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 299–305. Association for Computational Linguistics. 1, 8

Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. 2019. Aspect-based sentiment analysis using BERT. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics, NoDaLiDa 2019, Turku, Finland, September 30 - October 2, 2019*, pages 187–196. Linköping University Electronic Press. 8

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear. 5

Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. 2008. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, volume 21, pages 689–696. Curran Associates, Inc. 8

Yan Cathy Hua, Paul Denny, Katerina Taskova, and Jörg Wicker. 2023. A systematic review of aspect-based sentiment analysis (ABSA): domains, methods, and trends. *CoRR*, abs/2311.10777. 9

Dominik Janzing. 2007. On causally asymmetric versions of occam's razor and their relation to thermodynamics. *arXiv preprint arXiv:0708.3411*. 8

Dominik Janzing. 2019. The cause-effect problem: Motivation, ideas, and popular misconceptions. In Isabelle Guyon, Alexander R. Statnikov, and Berna Bakir Batu, editors, *Cause Effect Pairs in Machine Learning*, pages 3–26. Springer. 3, 8

Dominik Janzing, Rafael Chaves, and Bernhard Schölkopf. 2016. Algorithmic independence of initial condition and dynamical law in thermodynamics and

causal inference. *New Journal of Physics*, 18(9):093052. 8

Dominik Janzing, Joris M. Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniusis, Bastian Steudel, and Bernhard Schölkopf. 2012. Information-geometric approach to inferring causal directions. *Artif. Intell.*, 182-183:1–31. 8

Dominik Janzing and Bernhard Schölkopf. 2010. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194. 3, 8

Zhijing Jin, Julius von Kügelgen, Jingwei Ni, Tejas Vaidhya, Ayush Kaushal, Mrinmaya Sachan, and Bernhard Schoelkopf. 2021. Causal direction of data collection matters: Implications of causal and anticausal learning for NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9499–9513, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. 1, 2, 3, 8, 9

Daniel Kahneman. 2011. Thinking, fast and slow. *Farrar, Straus and Giroux*. 1, 2, 4, 8

Daniel Kahneman and Shane Frederick. 2002. *Representativeness revisited: Attribute substitution in intuitive judgment*, pages 49–81. Cambridge University Press. 4

Daniel Kahneman, Barbara L Fredrickson, Charles A Schreiber, and Donald A Redelmeier. 1993. When more pain is preferred to less: Adding a better end. *Psychological science*, 4(6):401–405. 2, 4, 8

Karim S Kassam and Wendy Berry Mendes. 2013. The effects of measuring emotion: Physiological reactions to emotional situations depend on whether someone is asking. *PloS one*, 8(6):e64959. 1, 8

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics. 1, 2, 4, 5, 8

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751. ACL. 8

Joseph E LeDoux. 1998. *The emotional brain: The mysterious underpinnings of emotional life*. Simon and Schuster. 4

Zhiheng Lyu, Zhijing Jin, Rada Mihalcea, Mrinmaya Sachan, and Bernhard Schölkopf. 2022. Can large language models distinguish cause from effect? In *UAI 2022 Workshop on Causal Representation Learning*. 9

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. 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. 2

Sergio Hernan Garrido Mejia, Elke Kirschbaum, and Dominik Janzing. 2022. Obtaining causal information by merging datasets with MAXENT. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, volume 151 of *Proceedings of Machine Learning Research*, pages 581–603. PMLR. 8

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *arXiv preprint arXiv:2202.05262*. 7, 14

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Noisy channel language model prompting for few-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5316–5330, Dublin, Ireland. Association for Computational Linguistics. 8

Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. 2014. Distinguishing cause from effect using observational data: Methods and benchmarks. *CoRR*, abs/1412.3773. 8

Rodrigo Moraes, JoãO Francisco Valiati, and Wilson P GaviãO Neto. 2013. Document-level sentiment classification: An empirical comparison between svm and ann. *Expert Systems with Applications*, 40(2):621–633. 8

Tetsuya Nasukawa and Jeonghee Yi. 2003. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 70–77. ACM. 1, 8

Jingwei Ni, Zhijing Jin, Markus Freitag, Mrinmaya Sachan, and Bernhard Schölkopf. 2022. Original or translated? A causal analysis of the impact of translationese on machine translation performance. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5303–5320, Seattle, United States. Association for Computational Linguistics. 3

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774. 6

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155. 6

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language*

Processing, EMNLP 2002, Philadelphia, PA, USA, July 6-7, 2002, pages 79–86. 8

Judea Pearl. 2001. Direct and indirect effects. In *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, University of Washington, Seattle, Washington, USA, August 2-5, 2001*, pages 411–420. Morgan Kaufmann. 7

Judea Pearl et al. 2000. *Causality: Models, reasoning and inference*. Cambridge University Press. 3

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2010. Identifying cause and effect on discrete data using additive noise models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, volume 9 of *JMLR Proceedings*, pages 597–604. JMLR.org. 8

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: Foundations and learning algorithms*. The MIT Press. 3

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*, pages 27–35. The Association for Computer Linguistics. 9

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2023. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Trans. Affect. Comput.*, 14(1):108–132. 1, 5

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8). 6

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67. 8

Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Sci.*, 5(1):31. 4, 15

Peter Salovey and John D Mayer. 2004. *Emotional intelligence*. Dude publishing. 1, 8

Ajay B Satpute, Jocelyn Shu, Jochen Weber, Mathieu Roy, and Kevin N Ochsner. 2013. The functional neural architecture of self-reports of affective experience. *Biological psychiatry*, 73(7):631–638. 1, 8

Timo Schick and Hinrich Schütze. 2022. True few-shot learning with Prompts—A real-world perspective. *Transactions of the Association for Computational Linguistics*, 10:716–731. 8

Bernhard Schölkopf. 2022. Causality for machine learning. In Hector Geffner, Rina Dechter, and Joseph Y. Halpern, editors, *Probabilistic and Causal Inference: The Works of Judea Pearl*, volume 36 of *ACM Books*, pages 765–804. ACM. 2, 3

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris M. Mooij. 2012. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress. 3

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Towards causal representation learning. *CoRR*, abs/2102.11107. 2

Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. 2018. Towards the first adversarially robust neural network model on mnist. *arXiv preprint arXiv:1805.09190*. 3

Naji Shajarisales, Dominik Janzing, Bernhard Schölkopf, and Michel Besserve. 2015. Telling cause from effect in deterministic linear dynamical systems. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 285–294. JMLR.org. 8

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642. 8

Peter Spirtes and Kun Zhang. 2016. Causal discovery and inference: Concepts and recent methodological advances. In *Applied informatics*, volume 3, pages 1–28. SpringerOpen. 3

Songbo Tan, Xueqi Cheng, Yuefen Wang, and Hongbo Xu. 2009. Adapting naive bayes to domain adaptation for sentiment analysis. In *Advances in Information Retrieval, 31th European Conference on IR Research, ECIR 2009, Toulouse, France, April 6-9, 2009. Proceedings*, volume 5478 of *Lecture Notes in Computer Science*, pages 337–349. Springer. 8

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca. 6

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971. 6

Ciprian-Octavian Truică, Elena-Simona Apostol, Maria-Luiza Șerban, and Adrian Paschke. 2021. Topic-based document-level sentiment analysis using contextual cues. *Mathematics*, 9(21). 1

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 417–424. ACL. 1, 8

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, 185(4157):1124–1131. 4

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual invariance to spurious correlations in text classification. *Advances in Neural Information Processing Systems*, 34. 9

Pranav Venkit, Mukund Srinath, Sanjana Gautam, Saranya Venkatraman, Vipul Gupta, Rebecca Passonneau, and Shomir Wilson. 2023. The sentiment problem: A critical survey towards deconstructing sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13743–13763, Singapore. Association for Computational Linguistics. 1

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias. 7

Janyce M Wiebe. 1994. Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287. 1, 8

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. 6

Xiaoyu Xing, Zhijing Jin, Di Jin, Bingning Wang, Qi Zhang, and Xuanjing Huang. 2020. Tasty burgers, soggy fries: Probing aspect robustness in aspect-based sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3594–3605, Online. Association for Computational Linguistics. 8, 9

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237. 5, 8

Robert B Zajonc. 1980. Feeling and thinking: Preferences need no inferences. *American psychologist*, 35(2):151. 4

Kun Zhang and Aapo Hyvärinen. 2009. Causality discovery with additive disturbances: An information-theoretical perspective. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part II 20*, pages 570–585. Springer. 3, 8

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28. 1, 2, 4, 5, 7, 8

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. 14

A Implementation Details

A.1 Model Details

Using Closed-Weight Models For the use of GPT model series, we use the OpenAI API,[3] with a text generation temperature of 0. We spent around 400 USD across around 20-30K single API calls.

Using Open-Weight Models For reproducibility, we set the generation temperature to 0 for all the models used in our work. For the open-weight models, GPT2-XL, LLaMa-7B and Alpaca-7B, it took around 24 hours on 4 GPUs RTX 2080 to generate their predictions on 1K data points for the 5 paraphrases of the causally-neutral prompt (denoted as C0), and on 500 data points for the 5 paraphrases of the C1 prompt, and 5 paraphrases of the C2 prompt. The causal tracing experiments with LLaMa-7B and Alpaca-7B on 100 data points took around 24 hours each using one GPU V100.

A.2 Sentence Sentiment Score Calculation

We apply the inverse sigmoid function to convert sentiment score probabilities from the range [0,1] to $[-\infty, \infty]$.

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) , \qquad (9)$$

where p is the probability of the positive class.

In practice, we cap the output of the logit function to the range $[-10, 10]$ as the sentence sentiment score. Namely,

$$\text{score}(p) = \max(-10, \min(10, \log(\text{logit}(p)))) . \qquad (10)$$

In this way, our scores fall between -10 and 10, corresponding approximately to probabilities of the positive label between 0.0001 and 0.9999.

To map our 5-class scores to the $[-10, 10]$ range, we assign the class labels as follows: -10 corresponds to the label 1, -5 to the label 2, 0 to the label 3, 5 to the label 4, and 10 to the label 5.

A.3 Implementation Details for Causal Tracing

We introduce the workings of the causal tracing method (Meng et al., 2022) as follows. First, we compute the hidden states of the residual stream of LLaMa-7B's layers for two inputs, (1) the original

---

[3] https://openai.com/api/

input: the prompt+review, and (2) the corrupted input: prompt + a corrupted version of the review by adding random noise immediately after the token embeddings. Then, we restore one by one the clean state of the residual stream into the corrupted version and measure the effect of the clean state on the probability of the originally predicted token for each token sequence and layer position.

Since this process is highly time-consuming, taking around 12 hours for 50 samples even using the smallest LLaMa model with 7B parameters, we do a case study on the 7B LLaMa and Alpaca using 100 random samples from the 1K test set. For these experiments, we follow the idea of APE (Zhou et al., 2023) to use the best-performing prompts on the 1k test set for C1 and C2.

A.4 Prompts

A.4.1 Prompts to Get Paraphrases

Since we need to report the average performance across five paraphrases of the same prompt, for each original prompt, we call GPT to generate the four paraphrases.

Below is the prompt that we used for this paraphrase generation process:

> You are an expert in prompt engineering for large language models (LLMs). And you are also a native English speaker who writes fluent and grammatically correct text.
>
> Given the following prompt for NLP sentiment analysis, you provide four alternative prompts.
>
> ####### Original Prompt #######
> [Our original prompt]
>
> ####### Alternative Prompt 1 #######
>
> ... (Then, we let the model to generate all the way to "Alternative Prompt 4".)

We queried the GPT-4 model with temperature 0 on June 8, 2023.

A.4.2 Neutral Prompt

In addition to the standard prompt to query LLMs in the main paper, we show its four paraphrases in Table 6.

Prompt Design
As a proficient data annotator in natural language processing (NLP), your responsibility is to determine the sentiment of the given review text. Please assign a sentiment value from "1" (very negative) to "5" (very positive). Review Text: "[review]" Sentiment Score:
As a skilled data annotator in the field of natural language processing (NLP), your task is to evaluate the sentiment of the given review text. Please classify the sentiment using a scale from "1" (highly negative) to "5" (highly positive). Review Text: "[review]" Sentiment Rating:
As an expert data annotator for NLP tasks, you are required to assess the sentiment of the provided review text. Kindly rate the sentiment on a scale of "1" (extremely negative) to "5" (extremely positive). Review Text: "[review]" Sentiment Score:
As a proficient data annotator in natural language processing (NLP), your responsibility is to determine the sentiment of the given review text. Please assign a sentiment value from "1" (very negative) to "5" (very positive). Review Text: "[review]" Sentiment Assesment:

Table 6: Four additional paraphrases of the neutral prompt (C0) generated with GPT-4.

A.4.3 Causal Prompts

In addition to the standard C1 and C2 prompts in the main paper, we show the four paraphrases for each of them in Tables 7 and 8, respectively.

B Additional Experimental Results

B.1 Few-Shot Results

For reproducibility and controllability, we use the zero-shot prompting setting across the experiments in the main paper, to avoid randomness in few-shot prompting according to which examples are selected as the few shots, and the order of the examples.

As a supplementary information in case this is of some readers' interest, we provide the few-shot prompting results in Tables 9 and 10.

B.2 λ_1-λ_2 Distribution Plot

To provide a clear understanding of the distributions of λ_1 and λ_2, we include their density plots of the causal processes C1 and C2 in Figure 3. The mean values of λ_1 and λ_2 for each group are in Table 11.

Further, we performed the Mann-Whitney U rank test to determine if the underlying distributions of

Prompt Design
As a customer sharing my experience, I crafted the following review: "[review]" Taking into account the details of my experience, I chose a star rating from the available options of "1","2","3", "4", or "5". My ultimate rating is:
As a client providing my opinion, I penned down the subsequent evaluation: "[review]" Upon thorough reflection of my encounter, I picked a star rating among the choices of "1","2", "3", "4", or "5". My conclusive rating stands at:
As a patron expressing my thoughts, I drafted the ensuing commentary: "[review]" After meticulously assessing my experience, I opted for a star rating from the range of "1","2","3", "4", or "5". My definitive rating turned out to be:
As a consumer conveying my perspective, I authored the following assessment: "[review]" By carefully weighing the aspects of my interaction, I determined a star rating from the possibilities of "1","2", "3", "4", or "5". My final verdict on the rating is:

Table 7: Four additional paraphrases of the causal prompt C1 generated with GPT-4.

λ_1 and λ_2 for groups C1 and C2 are the same. The results are as follows:

- For λ_1, the p-value is 8.4572×10^{-71}, leading us to reject the null hypothesis that the two groups come from the same distribution.
- For λ_2, the p-value is 1.36138×10^{-11}, also leading us to reject the null hypothesis that the distributions are the same.

These statistical results indicate significant differences between the distributions of λ_1 and λ_2 across the causal process groups, which indicate distinct underlying characteristics in the sentiment dynamics of the two groups.

C Emotion Arc Clustering

We analyze the emotional arc patterns of Yelp reviews. Reagan et al. (2016) identified 6 basic emotional arc shapes in stories. However, reviews are usually shorter and therefore present fewer variations. We take each sentence of the review and predict its sentiment. Then we divide the review into ten bins and compute an average sentence sentiment for each decile to make reviews with different lengths comparable. Reviews shorter than 10 sentences generate null values for some deciles which we fill with the information of the next decile. In Figure 5, we illustrate the 4 clusters found, they have the following characteristics:

Positive + Early Rise: This cluster primarily comprises highly positive reviews, where customers

Prompt Design
As a customer sharing my experience, I first chose a star rating from the available choices of "1","2", "3", "4", or "5", and subsequently elaborated on my decision with the following statement: "`[review]`" The review elucidates the reasoning behind my assigned rating of
As a client providing my opinion, I initially picked a star rating from the range of "1" to "5", and then proceeded to justify my selection with the following commentary: "`[review]`" The review sheds light on the rationale for my given rating of
As a patron expressing my thoughts, I started by selecting a star rating from the scale of "1" to "5", and then offered an explanation for my choice in the following review text: "`[review]`" The review expounds on the basis for my designated rating of
As a consumer conveying my perspective, I began by opting for a star rating within the "1" to "5" spectrum, and then detailed my reasoning in the subsequent review passage: "`[review]`" The review delineates the grounds for my conferred rating of

Table 8: Four additional paraphrases of the causal prompt C2 generated with GPT-4.

		Random	GPT-3 Few-Shot
F1	Overall	19.82 $_{\pm2.07}$	63.35 $_{\pm0.80}$
	C1 Subset	21.36 $_{\pm2.26}$	54.44 $_{\pm1.24}$
	C2 Subset	20.43 $_{\pm2.95}$	75.65 $_{\pm0.45}$
Accuracy	Overall	19.78 $_{\pm2.07}$	64.14 $_{\pm0.86}$
	C1 Subset	20.61 $_{\pm2.23}$	54.22 $_{\pm1.28}$
	C2 Subset	18.86 $_{\pm2.78}$	75.18 $_{\pm0.53}$

Table 9: Few-shot performance of the standard SA prompts on Yelp-5. We use five paraphrases for the prompt, and report the average performance with the standard deviation.

		Random	GPT-3 Few-Shot
F1	Data=C1, Prompt=C1	20.47 $_{\pm2.47}$	49.18 $_{\pm0.76}$
	Data=C1, Prompt=C2	20.26 $_{\pm2.31}$	52.79 $_{\pm2.64}$
	Data=C2, Prompt=C1	22.35 $_{\pm3.02}$	80.46 $_{\pm1.29}$
	Data=C2, Prompt=C2	20.35 $_{\pm2.18}$	75.88 $_{\pm1.86}$
Acc	Data=C1, Prompt=C1	19.60 $_{\pm2.67}$	50.12 $_{\pm0.71}$
	Data=C1, Prompt=C2	19.68 $_{\pm2.46}$	53.83 $_{\pm2.46}$
	Data=C2, Prompt=C1	20.97 $_{\pm3.19}$	81.21 $_{\pm1.17}$
	Data=C2, Prompt=C2	19.03 $_{\pm2.18}$	76.36 $_{\pm1.53}$

Table 10: Few-shot performance on Yelp using two different causal prompts on the two causal subsets. We use five paraphrases for each prompt, and report the mean performance with the standard deviation.
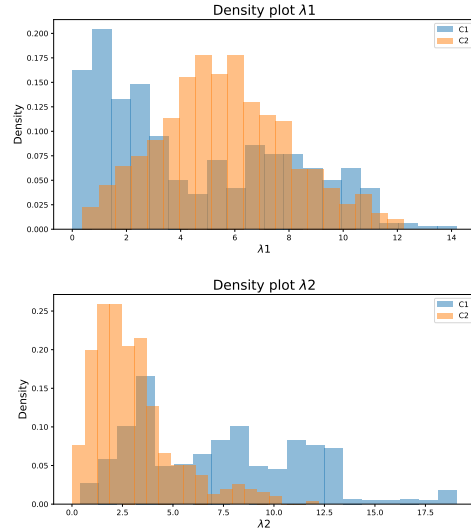

Figure 3: The λ_1-λ_2 density plots of C1 (above) and C2 (below).

express satisfaction and praise for their overall experience. Interestingly, 21.1% of these reviews begin with a negative first sentence, which often indicates initially low expectations or a negative first impression. However, despite the initial negativity, the reviews tend to turn positive as customers elaborate on their positive experiences.

Negative + Early Fall: This cluster mainly consists of predominantly negative reviews. Similarly to the Positive cluster, some reviews (28.14%) start with a sentence with the opposite sentiment, usually indicating high expectations followed by disappointment.

Rise: The main characteristic of this cluster is the positive ending of the review, despite the initial negativity observed in the first half, with an average sentiment of -1.63. An important fraction of the reviews in this cluster (52.49%) start with a positive comment as a summary, but then proceed

to highlight the negative aspects of the experience. Despite the initial criticisms, the reviews conclude with positive points, suggesting that the overall experience was still satisfactory.

Fall: In contrast to the previous cluster, the Fall cluster is characterized by a negative ending of the review, despite a generally positive first half with an average sentiment of 2.18. An important proportion (36%) of the reviews in this cluster begin with a negative comment as a summary, but then proceed to describe the positive aspects before eventually highlighting the negative ones. This cluster showcases a shift in sentiment from positive to negative, indicating a decline in satisfaction as the review progresses.

D Additional Interpretability by Shapley Values

We further analyze the effect of each part of the prompts on LLaMa's predictions. Using 50 reviews, we compute the shapley values of each to-

	C1	C2
$\mu(\lambda_1)$	4.48	5.62
$\mu(\lambda_2)$	7.31	3.02

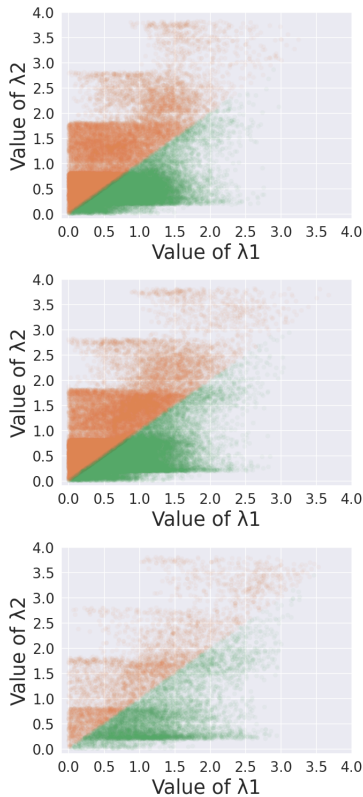Table 11: Mean values of the lambdas for C1 and C2.

Figure 4: The λ_1-λ_2 plot on Yelp-5 (left), Amazon (middle), and App Review (right). We draw the $y = x$ diagonal line, and the orange dots in the upper-left triangle represent the C1-dominant subset, and green dots in the lower-right triangle are the C2-dominant subset.
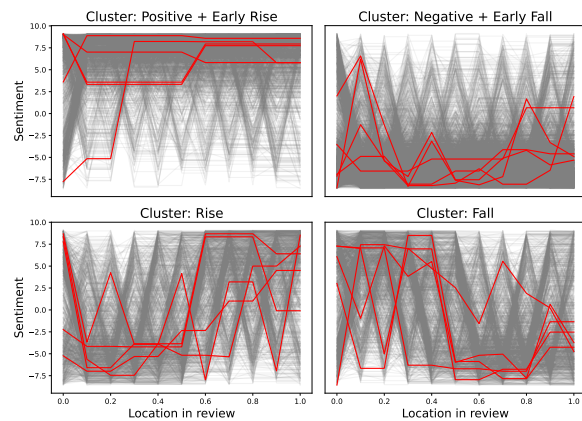

Figure 5: Four emotion arc clusters.

ken. In Figure 6 we observe that the tokens with the largest shapley values are the ones in the end, which is expected since they are the ones helping to form a grammatically correct sentence. To account for that, we subtracted the average shapley values computed for the other possible start rating answers. In Figure 7 we show the adjusted shapley values. We observe that the tokens in prompt C1 have a larger effect than the tokens in prompt C2. The words introducing the review have a positive effect on C2 but a negative one on C1. Whereas, the phrase "I chose a star rating" has a negative effect on C2 but a positive one on C1.
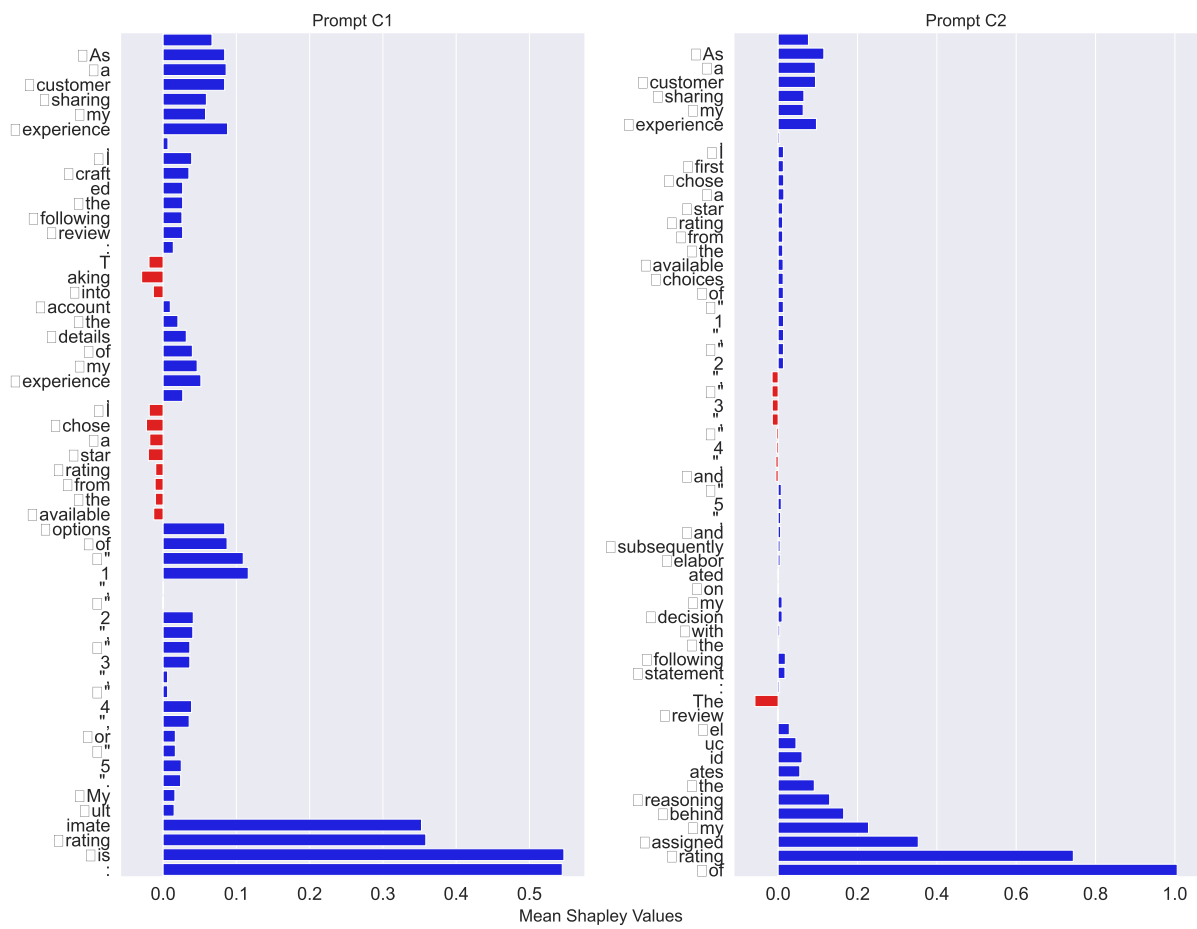
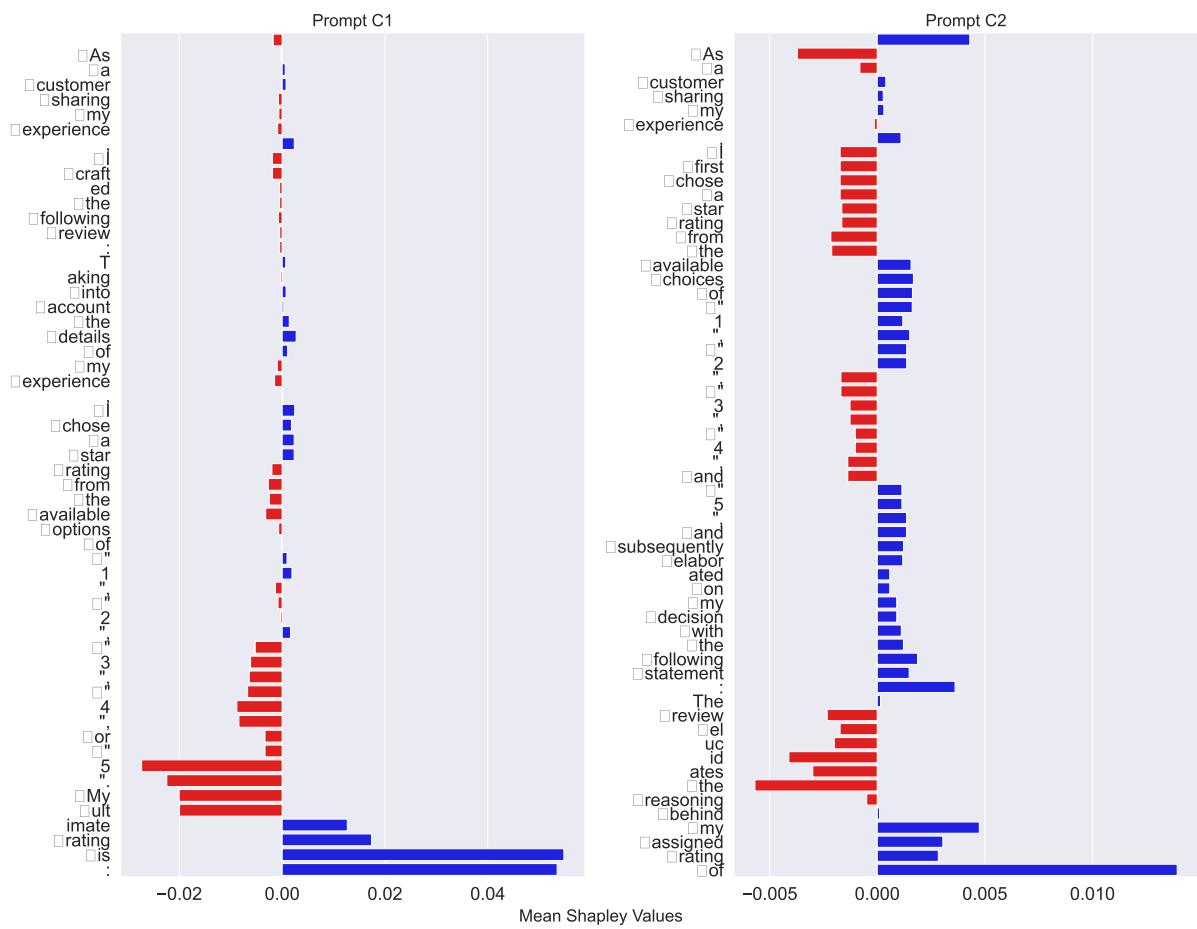Figure 6: Shapley values for the 2 types of prompt.

Figure 7: Adjusted shapley values for the 2 types of prompt.

Positive + Early Rise

Review: *Was there last Friday. Seats right in front if the stage. The show was good. The headliner, while a bit long, was good. Fantastic service from our waitresses. Will definitely go back.*

Review: *This is by far my favorite Panera location in the Pittsburgh area. Friendly, plenty of room to sit, and good quality food & coffee. Panera is a great place to hang out and read the news - they even have free WiFi! Try their toasted sandwiches, especially the chicken bacon dijon.*

Negative + Early Fall

Review: *Pass on this place, there are better restaurants mere feet away.*

The menu here is too large, which is a sure sign none of the food is going to be good. And, its not good. Some of the salads are alright, but its just not good food.

The service is friendly and prompt, but the beer is over priced. They do have a good selection though.

This place is open late if you need a bite to eat, but there are so much better options out there.

Review: *Wings are overpriced. And the quality of them are bad. They were tough and greasy. The staff are pleasant but then over all experience was too expensive for a sports bar.*

Rise

Review: *To be honest, I feel that this is one of the most overpriced restaurants in the entire city. The food is average to good, the place is beautiful with outdoor seating, but in my opinion the price is just not worth it. They have a really good happy hour, so I would definitely recommend going to that and maybe trying an appetizer or two.*

Review: *The first time I came here, I waited in line for 20 minutes. When it was my turn, I realized I left my wallet in the car. It hurt so bad, I didn't come back for a year.*

I can walk to this place from my house- which is dangerous because those biscuits are just OH SO DREAMY. I can't describe them. Just get some.

Do I feel guilty about noshing on fabulous Strawberry Napoleons and Jewish Pizza (kind of like a modified, yet TOTALLY delicious fruitcake bar) at 10:15am? Hecks, naw... But they do have quiche and some other breakfast-y items for those who prefer a more traditional approach to your stomach's opening ceremony.

Just go early :) They open at 10 on Saturdays. And bring cash...it's easier that way.

Fall

Review: *It's cheap, I'll say that, but otherwise it's bland food served by workers who mostly don't seem to notice they're working, and when they do, only respond snarkily. There are many better vegetarian and vegan options to choose from*

Review: *I do like my Mad Mex, however predictable and non-authentic it may be. The portion sizes are mammoth and I come away with a satisfied sense of regret. Their beer menu is happily extensive. Charging me $9 for chips and salsa is a bit of crime, wouldn't ya say though!?! I mean, c'mon! Our service has most times been lacking– a bit rushed and on the inattentive side. Also, why do you require your wait staff to not servestraws/ lemons/etc unless asked by cusotmers—weirdness-cut out these odd cost-cutting, anti-service friendly measures please*

Table 12: Example reviews for each emotion arc cluster.