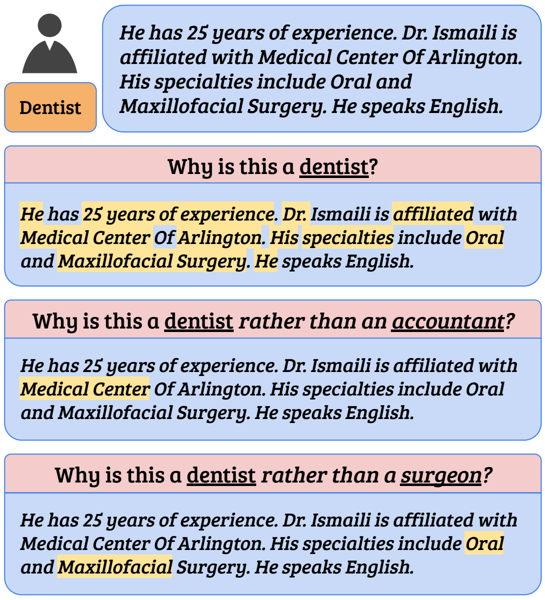
Contrastive Explanations for Model Interpretability

Alon Jacovi[♡*] Swabha Swayamdipta[♣] Shauli Ravfogel[♡♣]
Yanai Elazar[♡♣] Yejin Choi[◊♣] Yoav Goldberg[♡♣]
[♡]Bar Ilan University
[♣]Allen Institute for Artificial Intelligence
[◊]Paul G. Allen School of Computer Science and Engineering, University of Washington
alonjacovi@gmail.com
{swabhas,shaulir,yanaie,yoavg,yejinc}@allenai.org

Abstract

Contrastive explanations clarify why an event occurred in contrast to another. They are more inherently intuitive to humans to both produce and comprehend. We propose a methodology to produce contrastive explanations for classification models by modifying the representation to disregard non-contrastive information, and modifying model behavior to only be based on contrastive reasoning. Our method is based on projecting model representation to a latent space that captures only the features that are useful (to the model) to differentiate two potential decisions. We demonstrate the value of contrastive explanations by analyzing two different scenarios, using both high-level abstract concept attribution and low-level input token/span attribution, on two widely used text classification tasks. Specifically, we produce explanations for answering: for which label, *and against which alternative label*, is some aspect of the input useful? And which aspects of the input are useful for and against particular decisions? Overall, our findings shed light on the ability of label-contrastive explanations to provide a more accurate and finer-grained interpretability of a model's decision.



He has 25 years of experience. Dr. Ismaili is affiliated with Medical Center Of Arlington. His specialties include Oral and Maxillofacial Surgery. He speaks English.

Why is this a dentist?

He has 25 years of experience. Dr. Ismaili is affiliated with Medical Center Of Arlington. His specialties include Oral and Maxillofacial Surgery. He speaks English.

Why is this a dentist *rather than an accountant?*

He has 25 years of experience. Dr. Ismaili is affiliated with Medical Center Of Arlington. His specialties include Oral and Maxillofacial Surgery. He speaks English.

Why is this a dentist *rather than a surgeon?*

He has 25 years of experience. Dr. Ismaili is affiliated with Medical Center Of Arlington. His specialties include Oral and Maxillofacial Surgery. He speaks English.

Figure 1: Illustrative example with explanations as textual highlights (yellow), showing biodata and occupation from BIOS (De-Arteaga et al., 2019). Explanations without an explicit contrast (left) are potentially misaligned with human expectations of what is being explained, making them confusing for human interpretability. Contrastive explanations (middle; right) prune the space of all causal factors to intuitively relevant ones, allowing finer-grained understanding, and can vary substantially based on the contrasted decision.

1 Introduction

Explanations in machine learning models attempt to discover the causal factors leading to the model decisions. Methods for producing model explanations often seek all causal factors at once—making them difficult to comprehend—or organize them via heuristics, such as gradient saliency (Simonyan et al., 2013; Li et al., 2016). However, it remains unclear what makes a particular collection of causal factors a good explanation. Should we seek all factors behind a decision, or perhaps report causes that are most *important* to the decision (via some measure)? In this work, we discuss how existing approaches are not only too coarse-grained, but also potentially misleading.

Studies in social science have established that human explanations, on the other hand, are typically "contrastive" (Miller, 2019): they rely on the causal factors that explain why an event occurred *instead of* an alternative event (Lipton, 1990). Such explanations reduce cognitive load both for the explainer and the explainee, by pruning the space of all causal factors, which promotes easier communication between the explainer and explainee, as illustrated in Figure 1.

[*] Work done during an internship at the Allen Institute for Artificial Intelligence.

This reduction is deeply relevant to model explanations: since explanations of opaque ML and NLP models are *approximations* of complex statistical processes, human interpretability of these decisions already involves making subjective decisions about which aspects of the decision to explain. Additionally, humans inherently assume some contrast decision when deciphering these explanations (whether we intend for this or not) (Hilton and Slugoski, 1986; Hesslow, 1988). We seek to design explanations that are explicitly contrastive, in order to reveal fine-grained aspects of model decisions in a comprehensible manner, and are more accurate to how human observers comprehend them.

In this work, we propose an explicitly contrastive methodology for analyzing model behavior. We begin with a background on contrastive explanations as humans understand them, and with a motivation and structure for deriving such explanations from machine learning models (§2). Why do humans comprehend explanations contrastively? And from understanding this question, what can we learn about what contrastive explanations of models should satisfy?

We then propose a framework for deriving contrastive explanations, and a modification to be performed on trained neural classifiers (§3). The framework operates on the input and representation spaces, and produces a contrastive latent representation—and therefore contrastive model reasoning—of a given model. This is accomplished by projecting the latent representations of the model (on given examples) to the space that minimally separates two decisions in the model. Our framework is robust, useful and flexible; it can be adapted to existing explanation methods to produce contrastive explanations.

We demonstrate this flexibility by describing an analysis of two well-studied text classification datasets: BIOS (De-Arteaga et al., 2019) (§4) and MultiNLI (Williams et al., 2018) (§5), and models trained on them.[1] Specifically, we showcase experiments where:

1. We assess whether textual highlights (Lei et al., 2016), or discrete portions of the input, are contrastively relevant. For example, we find that *pronouns* and *personal names* are relevant to a model trained on the BIOS dataset to classify occupations of persons—in deciding that an example describes a *paralegal* rather than an *attorney*.

2. We identify whether a particular *concept* (abstract semantic feature in the input) is contrastive. For example, we find that the *overlap* concept (whether the hypothesis is lexically overlapped with the premise) in MultiNLI is useful for the model in deciding that an example is *entailment* rather than *neutral*.

3. We identify which highlights are causally linked to the contrastive decision of one decision over another, via input-ablation behavioral analysis, demonstrated on both BIOS and MultiNLI.

4. We discover which alternative decisions are most and least dominant in the model's reasoning when predicting a particular label.

2 Contrastive Explanations

Explanations can be considered as answers to the question: "why P?" where P, the **fact**, is the event to be explained. For example:

Why did you decide to hire *[Person X]*?

The explanation behind P (the hiring decision) comprises the causal chain of events that led to P. In the above case, the answer may cite that *Person X* has a relevant degree, and some professional experience.

However, explaining the *complete* causal chain is both burdensome to the explainer and cognitively demanding to the explainee (Hilton and Slugoski, 1986; Hesslow, 1988). For instance, the above event was distantly caused by *Person X*'s birth—yet the explainer will likely omit this factor from the explanation. Indeed, this omission simplifies the explanation, reducing cognitive load. But which factors should be omitted, and which should not?

Contrastive explanations theory clarifies this question: humans inherently position explanations as an answer to the question: "why P, rather than Q?" (Hilton, 1988), where Q (the **foil**[2]) is some

[1] Implementation available at `https://github.com/allenai/contrastive-explanations`.

[2] Q is generally termed as the *contrast fact*, of which it can either be a 'foil' (result of a counterfactual context) or a 'surrogate' (result of a bifactual context) (Miller, 2020)—where a counterfactual context refers to an imaginary (e.g., synthetic) one designed specifically in counter to some fact, and a bifactual context refers to a naturally-occurring context that results in the foil. In this work, we consider only counterfactual contexts, hence we use the term foil to refer to possible contrast facts.

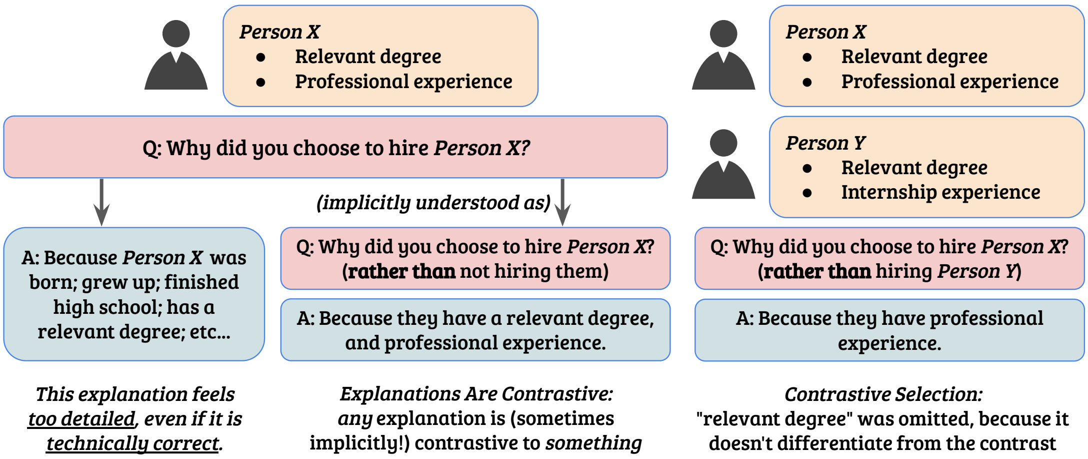

Person X
Relevant degree
Professional experience
Q: Why did you choose to hire Person X?
(implicitly understood as)
A: Because Person X was born; grew up; finished high school; has a relevant degree; etc...
Q: Why did you choose to hire Person X? (rather than not hiring them)
A: Because they have a relevant degree, and professional experience.
Person X
Relevant degree
Professional experience
Person Y
Relevant degree
Internship experience
Q: Why did you choose to hire Person X? (rather than hiring Person Y)
A: Because they have professional experience.
This explanation feels too detailed, even if it is technically correct.
Explanations Are Contrastive: any explanation is (sometimes implicitly!) contrastive to something
Contrastive Selection: "relevant degree" was omitted, because it doesn't differentiate from the contrast

Figure 2: A simple demonstration of contrastive explanation theory (§2).

alternative event. Therefore, the explanation to the hiring decision may be answering:

> Why did you decide to hire *[Person X]*, rather than not hiring them?

Since the decision not to hire the candidate was also affected by their birth, by explaining contrastively, the explanation can be simplified by omitting this factor; this illustrate **contrastive causal attribution** in the explanation process.

Similarly, given a different foil:

> Why did you decide to hire *[Person X]*, rather than hiring *[Person Y]*?

Suppose that *Person Y* has the same attribute of a "relevant degree" as *Person X*—then the explainer may feel that it is unnecessary to attribute this factor to their decision as well—only mentioning the professional experience. Figure 2 demonstrates contrastive explanations.

Implications for model explanations. Explanations in machine learning are aimed to discover factors which can be causally[3] attributed to the model's decision—for example, to investigate whether the decision is being made for the right reasons (Ross et al., 2017; McCoy et al., 2019), or to allow path of recourse to parties affected by the decision (Ustun et al., 2019). Therefore, model explanations stand to benefit significantly from explicit contrastive structure, in two ways:

1. **Ease of explaining:** Model decisions are complex and noisy statistical processes. Therefore, complete explanations of these decisions is difficult (Jacovi and Goldberg, 2020b). Reducing the burden of what to explain via contrastive selection will naturally moderate this process.

2. **Ease of understanding:** Humans inherently comprehend explanations contrastively. If a foil is not explicitly mentioned (as with most current explanation methods), it is implicitly assumed by the explainee in some, likely unintended, form. This assumption may not match the foil which the explanation implicitly entails (Kumar et al., 2020), leading to misalignment between the explanation's meaning and how it is understood. Thus we should strive to design explicitly contrastive explanations.

3 A Framework for Contrastive Explanations

We propose a framework for contrastive attribution of candidate causal factors, based on a background for understanding contrastive explanations (§3.1). We start with a discussion of causal attribution which yields model explanations via interventions; such explanations are not explicitly designed to be contrastive (§3.2). Next, we describe our framework for contrastive attribution via latent representations obtained using a powerful and flexible approach of row-space projections (§3.3).

3.1 Explanation Structure

Prior to describing how to derive contrastive explanations, we must first understand what form they

[3] Many current explanation methods in NLP involve measures of correlation, rather than causation (Feder et al., 2020). Whether they truly concern causation is a question of *faithfulness* (Ribeiro et al., 2016; Wiegreffe and Pinter, 2019), and is detached from the discussion in this section.

should take. This involves defining the following properties:

Candidate factor space. Model explanations communicate causal factors behind a decision. Therefore, we must define a space of candidate causal factors, and then attribute some of them to the decision via some method (i.e. attributing causality to them). Typical candidate factor spaces include *portions of the input (highlights)* (Lei et al., 2016), abstract *concepts* in the input (Kim et al., 2017), or *influential examples* in the training set (Koh and Liang, 2017; Han et al., 2020) that led to the model's decision. We base our studies on explanations in the form of *highlights* and *concepts*, though our framework is agnostic to the candidate factor space, so long as it is possible to intervene on their presence (§3.2).

Event space. The event we attempt to explain (the fact) is the trained model's final decision, i.e., we attempt to explain why the class was assigned higher probability than the others. Similarly, the foils we consider are all other classes, *separately*, as they encompass the rest of the possible alternative events. Hence our event space is the union (as a discrete set) of all possible classes.[4]

Contrastive selection. Deriving contrastive explanations involves selecting a set of causal factors from the candidate factor space, with respect to a foil from the event space. We consider two possible settings for explaining by suspending either of the two variables, in order to select the other:

1. **Ranking factors (fixed foil):** given a causal factor space (such as highlights or concepts) and a foil, we rank each factor by how much they are contrastively useful to the model towards the fact and against the foil.

2. **Ranking foils (fixed factor):** given a causal factor in the explained fact, we rank the set of available foils by how much the said factor is contrastively used by the decision process between the fact and the foils.

Note that the above adopts a relaxed continuous and relative perspective on contrastive selection, as opposed to the discrete and binary perspective discussed in Section 2. We leave discussion of possible discretization of this process to future work.

3.2 Causal Attribution via Interventions

Given a candidate factor space, we seek to attribute factors with causality over the decision process. We adopt an interventionist approach to measure causal attribution: this involves attributing causality to a factor by intervening on this factor; this produces a counterfactual. Change in model behavior under this counterfactual indicates the importance of the intervened factor.

Most interventions we consider are amnesic (Elazar et al., 2020), i.e. use counterfactuals which omit the candidate factors under consideration.[5] Explanations in terms of *highlights* and *concepts* require different intervention procedures. For *highlights*, we mask the relevant highlights, and consider models where masked data points are in distribution (Zintgraf et al., 2017; Kim et al., 2020), i.e. pre-trained masked language models, such as `RoBERTa` (Liu et al., 2019). For *concepts*, we employ an *amnesic operation* (Elazar et al., 2020) to probe for causality of the model. Where possible, we also produce *manually-annotated counterfactuals* via input-level modifications of existing dataset examples for concepts.

3.3 Contrastive Attribution via Row-Space Projection

While causal interventions yield the candidate factor space for explanations, we need to further prune this space for selecting contrastive explanations. Potential methods could include a second stage of interventions such as masking portions of the explanation or further amnesic operations in the concept space. However, such stage-2 interventions assume that the explanation from the first phase are correct; and that the contrastive factors are a proper subset of the causal factors discovered in the first phase of interventions.

Hence, we propose to measure model behavior contrastively, and produce contrastive explanations as latent representations, produced via a row-space projection of the input representation, in accordance with this measure of behavior: We establish contrastive behavior (§3.3.1) and describe the projection (§3.3.2) which allows the internal

[4] Therefore, we do not consider explanations which describe other aspects of model behavior: such as why a particular class was assigned a particular probability, or why the model predicted a particular class in contrast to a *combination* of other classes.

[5] In contrast to replacing them with other causal factors, instead; we discuss relative strengths and weaknesses of amnesic and non-amnesic interventions in §7.

representations of the model to only consider features which distinguish the fact and the foil.

3.3.1 Contrastive Behavior

Let $f(\mathbf{x})$ be a classification model that classifies a text input $\mathbf{x}$ to one of K classes $\mathcal{Y} = \{y_1, ..., y_K\}$. The model f commonly corresponds to an arbitrarily deep neural encoder $enc(\cdot)$ that transforms the input $\mathbf{x}$ into a vector $\vec{h}_\mathbf{x} \in \mathbb{R}^d$. This is followed by a final linear layer $\mathbf{W} \in \mathbb{R}^{d\times K}$, such that $f(\mathbf{x}) = \mathbf{W}enc(\mathbf{x}) = \mathbf{W}\vec{h}_\mathbf{x}$ yields the logits of the model over the K classes.

Let $y^* = \arg\max_{y\in\mathcal{Y}} f(\mathbf{x})$ be the class predicted by the model (the *fact*), y' be any alternative prediction one aims to focus on (the *foil*), and $\vec{p} := softmax(\mathbf{W}\vec{h}_\mathbf{x})$ the normalized model probabilities.[6] Finally, let $\vec{q} := softmax(\mathbf{W}\vec{h}_{\mathbf{x}'})$ be the model probabilities following some intervention which produces counterfactual $\mathbf{x}'$.

Then firstly, we measure contrastive model behavior as

$$\frac{p_{y^*}}{p_{y^*} + p_{y'}} - \frac{q_{y^*}}{q_{y^*} + q_{y'}}, \tag{1}$$

or the difference of normalized probabilities of the fact and foil. In other words, the preference of y^* over y' (the fact and foil, respectively) is measured as the difference in the logit assigned to y^* only in comparison y'—if p_{y^*} decreased ($p_{y^*} > q_{y^*}$) following the intervention, then $p_{y'}$ necessarily increased ($p_{y'} < q_{y'}$).

3.3.2 Contrastive Representations

Following the measurement in Equation 1, we can derive a contrastive latent representation of $\vec{h}_\mathbf{x}$ using *row space projection.*

The output of the model, $f(\mathbf{x})$ is linear in the latent encoding $\vec{h}_\mathbf{x}$. Let $\vec{\mathbf{w}}_i$ be the row in $\mathbf{W}$ corresponding to class i. The logits for classes y', y^* are given by the dot products $\vec{\mathbf{w}_{y'}}^T\vec{h}_\mathbf{x}$ and $\vec{\mathbf{w}_{y^*}}^T\vec{h}_\mathbf{x}$ respectively. The logits for each class are thus *unrelated* to any other row in the prediction matrix. These dot products are un-normalized projections of the representation $\vec{h}_\mathbf{x}$ on the directions $\vec{\mathbf{w}_{y^*}}, \vec{\mathbf{w}_{y'}} \in \mathbb{R}^d$. While the representation $\vec{h}_\mathbf{x}$ can be high-dimensional, only two directions (components) in this high-dimensional space are relevant for each contrastive decision.

Furthermore, we are interested in the prediction of one class over the other, as opposed to the logits. Hence, we can replace these two directions with a *single* contrastive direction $\vec{\mathbf{u}} \in \mathbb{R}^d$, by defining $\vec{\mathbf{u}} = \vec{\mathbf{w}_{y^*}} - \vec{\mathbf{w}_{y'}}$.

Given that the model favors y^* over y', i.e., $p_{y^*} > p_{y'}$, if and only if $\vec{\mathbf{u}}^T\vec{h}_\mathbf{x} > 0$, the projection of $\vec{h}_\mathbf{x}$ onto $\vec{\mathbf{u}}$ precisely derives the linear direction in $\mathbb{R}^d$ which the model uses to differentiate between classes y^* and y'. We refer to the *row-space* of $\vec{\mathbf{u}}$, represented as

$$\begin{aligned} Span(\vec{\mathbf{u}}) &= Span(\vec{\mathbf{w}_{y^*}} - \vec{\mathbf{w}_{y'}}) \\ &:= \alpha\vec{\mathbf{w}_{y'}} - \alpha\vec{\mathbf{w}_{y'}} \ . \ \forall\alpha \in \mathbb{R}, \end{aligned}$$

as the *contrastive space* for y^* and y'. We define the contrastive transformation $C(\vec{h}_\mathbf{x})_{y',y^*}$ to be the orthogonal projection onto this subspace:

$$C(\vec{h}_\mathbf{x})_{y',y^*} := \frac{uu^T}{u^Tu}\vec{h}_\mathbf{x} = \mathbf{P}_{\vec{u}}\vec{h}_\mathbf{x},$$

where $\mathbf{P}_{\vec{u}} := \frac{uu^T}{u^Tu}$ is a projection matrix onto $\vec{\mathbf{u}}$.

To give further motivation for the use of $\vec{\mathbf{u}}$, recall that all directions $\vec{v}$ orthogonal to $\vec{\mathbf{u}}$ form the *nullspace* of $\vec{\mathbf{u}}$ and satisfy $\vec{v}^T\vec{\mathbf{u}} = \vec{v}^T(w_{y^*} - w_{y'}) = \vec{0}$, or, equivalently, $\vec{v}^Tw_{y^*} = \vec{v}^Tw_{y'}$. Such directions support the classes y' and y^* to the same extent, and are thus contrastively irrelevant; we can discard those directions from $\vec{h}_\mathbf{x}$, without influencing the logits for y' and y^*.

The resulting representation $\vec{g} := C(\vec{h}_\mathbf{x})_{y',y^*}$ is a latent vector of the same dimensions as $\vec{h}_\mathbf{x}$. It captures (precisely) the latent features in $\vec{h}_\mathbf{x}$ which are used by the model to differentiate the fact from the foil. This provides a robust and versatile foundation from which to analyze model behavior contrastively in latent space.

4 Case Study I: Analyzing BIOS

We apply our framework to a multi-class classification task use-case on the BIOS dataset (De-Arteaga et al., 2019): a collection of biographies for individuals, containing descriptions of, or anecdotes from their professional careers, paired with their associated professions, as well their binary gender[7] (see Table 1).

The task involves classifying each biography text as a profession, from a total of 27 professions.[8] We

[6] Formally, the methods in this section do not require the fact to be the final model prediction, though we only consider this case in our experiments.

[7] The BIOS dataset treats the gender attribute as binary; this influences our studies. We acknowledge that this simplification cannot capture the entirety of gender and erases people who do not identify with these binaries.

[8] The original dataset contains 28 profession; we omit all biographies with the *model* profession from the entire dataset, as we found the annotations of this class to be noisy.

Biography	**Profession**	**Gender**
She also works as a Restitution Specialist while being the liaison to the Victim Compensation Board. Ms. Azevedo was named an OVSRS Outstanding Partner due to her dedication to providing critical information to staff so victims can obtain their court-ordered restitution while offenders can be held accountable. Ms. Azevedo always responds to staff inquires promptly and goes above and beyond to ensure victim's constitutional rights to restitution collection are being met.	paralegal	female
Ms. Wells was admitted to the California Bar in 2015, but she has worked for the firm since 2012. Ms. Wells is admitted to practice before all of the California state trial courts and the United States District Court for the Southern and Central Districts of California.	attorney	female
Peter also has substantial experience representing clients in government investigations, including criminal and regulatory investigations, and internal investigations conducted on behalf of clients.	attorney	male

Table 1: Examples from BIOS (De-Arteaga et al., 2019). The highlights (yellow) indicate information in each biography, such as names and pronouns for individuals, revealing demographic information. These highlights could be (unintentional, yet) potential factors explaining a model's prediction of the associated profession.

Fact (y^*)	Top 3		Bottom 3	
	Foil	$p_{y^*} - q_{y^*}$	Foil	$p_{y^*} - q_{y^*}$
paralegal	attorney	10.519	composer	0.019
	accountant	1.165	dj	0.021
	professor	0.387	surgeon	0.026
physician	professor	0.622	nurse	-0.315
	surgeon	0.406	chiropractor	-0.132
	psychologist	0.185	paralegal	-0.013
attorney	professor	1.292	paralegal	-0.223
	journalist	0.545	composer	0.018
	teacher	0.420	rapper	0.022
nurse	physician	3.531	paralegal	0.067
	surgeon	1.779	composer	0.105
	chiropractor	1.762	interior designer	0.117
yoga teacher	teacher	1.495	attorney	-0.128
	psychologist	0.399	nurse	-0.099
	photographer	0.190	accountant	-0.060
rapper	dj	1.878	paralegal	0.130
	composer	1.844	dietitian	0.177
	poet	1.335	interior designer	0.272
interior designer	architect	4.095	composer	0.081
	photographer	2.946	chiropractor	0.093
	journalist	2.046	paralegal	0.105

Table 2: The top-3 and bottom-3 contrastive foils for the impact of pronouns and names in BIOS for predicting the fact, on six classes. *Top 3/positive* means that pronouns and names were useful for the fact against the foil. *Bottom 3/negative* means that they were useful for the foil, instead.

analyze a model based on `RoBERTa-Large` (Liu et al., 2019) fine-tuned on the BIOS training set, with test-set performance of 87.52%.

4.1 The Role of Pronouns and Names (Demographic Attributes)

BIOS examples, as short anecdotes of a person's career, may contain information which is not semantically related to solving the task, but are nevertheless unintended useful signals to trained models. Table 1 contains examples of *paralegal* and *attorney* professions, where we observe the presence of such signals, such as the person's pronouns and personal name. These signals can reveal demographic information of the person's gender, country of origin, ethnicity, and so on (e.g., *Azevedo* is a common Portuguese surname).

Pronouns and names are therefore correlated with demographic attributes. These attributes in turn are correlated with specific professions in the dataset. De-Arteaga et al. (2019); Romanov et al. (2019) have previously shown that this information is typically leveraged by BIOS models—therefore *pronouns and names could explain models decisions. But how are they used* ***contrastively****?*

For example, female pronouns are correlated with the *paralegal* profession, as the people of this profession are overwhelmingly women in the BIOS training set (roughly 90%).

The pronouns and names in the input define a single highlight which we can then use in the *foil ranking* setup. We ask the question: *which classes does the model use this highlight as evidence against when making its decision?*

Table 2 shows results for multiple facts, and the top-3 and bottom-3 foils for each class on our metric of decrease in the normalized fact logit (§3.3), as average over the BIOS dev-set. A pair in this table with a positive measurement (> 0) means that the BIOS model, on average when predicting the fact, uses personal information encoded in the pronouns and the name of the person as evidence *for* the fact *against* the foil, to the degree of $p_{y^*} - q_{y^*}$. A pair with negative measurement means the vice-versa applies.

For example, on *paralegal*-predicted examples, *attorney* is strongly measured as a relevant foil. This implies that the model is often using pronouns and names as evidence for predicting *paralegal* rather than *attorney*. Such personal attributes are also useful to the model for predicting *attorney* in-

fact (% males)	**foil** (% males)	$p_{y^*} - q_{y^*}$	**sign(*cos*)**
paralegal (9%)	attorney (62%)	0.804	$-$
professor (55%)	teacher (41%)	0.225	$+$
accountant (63%)	psychologist (37%)	0.108	$+$
nurse (9%)	physician (47%)	0.084	$-$
teacher (41%)	poet (53%)	0.072	$-$

Table 3: The top-5 pairs of classes in the BIOS dataset for which the gender concept was measured to have the largest contrastive effect. The top foil is reported for each fact among the top-5 pairs. The sign of cosine similarity specifies whether the concept of "male" is evidence of the fact ($+$) or the foil ($-$), while the decrease in probability specifies the degree that the concept affected the decision process.

stead of *professor*, but are negative evidence against the *paralegal* foil. Similarly, the highlights evidence *nurse* over *physician*.

This illustrates the power that a contrastive perspective gives us in understanding model behavior: the model is leveraging demographic attributes as evidence for decisions between semantically similar classes, which are demographically different in some way.

4.2 The Role of Gender

The BIOS dataset provides binary gender concept labels for each instance (Table 1). We will use this concept to demonstrate our methodology for *conceptual* factors in the input. Previous work (Ravfogel et al., 2020) has shown that neural BIOS models indeed make use of the gender concept.

In order to intervene on the presence of the concept, we utilize *amnesic probing* (Elazar et al., 2020).[9] This method involves removing the gender information from $h_{\mathbf{x}}$ using *iterative null-space projection* (Ravfogel et al., 2020).

We measure the behavioral change for all pairs of classes, and report the top-5 pairs in Table 3. The top-scoring pairs trend to be semantically similar and are opposites in their gender majority (such as *paralegal* and *attorney* or *nurse* and *physician*), indicating that the model is leveraging the gender concept to differentiate between otherwise semantically-similar classes.

We additionally report *sign*$(\frac{<\vec{u},\vec{r}>}{\|\vec{u}\|\|\vec{r}\|})$, or the sign of the cosine similarity between $\vec{u}, \vec{r}$, where $\vec{u} = \vec{w_{y^*}} - \vec{w_{y'}}$ and $\vec{r}$ is the final concept probe trained in the amnesic probing procedure. This measure assigns a direction to the concept (in this case, the concept of the person being male) as a concept supportive towards the fact or towards the foil. The results align with intuition: for example, the concept of male is supportive of *attorney* over *paralegal*, and *accountant* over *psychologist*, as those are male-majority professions in the BIOS dataset.

4.3 Ranking Highlights by Contrastive Power

Sections 4.1, 4.2 have so far focused on explaining which foil a factor is contrastive against (*ranking foils*). In this section, we derive which which factors are contrastive to a given foil (*ranking factors*).

For simplicity, for our candidate factor space we consider all word unigram and bigram highlights. We assume a relevant foil is given, which can be any alternative class to the fact. We derive the model decision after intervening on each candidate, and measure the change in behavior.

We apply this technique towards understanding model errors, by selecting examples of model mistakes and assigning the foil to be the gold label. We show examples in Table 4 with the top-ranking highlight for answering the question: *which unigram or bigram was most relevant for the model in making its prediction rather than the gold label?*

For example, in the *surgeon* example (for which the model mistakenly predicted *physician*), the model is generally most affected by the bigram "top medicine". However, this is not a particularly useful feature to favor *physician* rather than *surgeon*, since surgeons also entail medicine studies; when we repeat the experiment in contrast to *surgeon*, the top highlight changes to "patients died", indicating that this bigram is a better differentiator for those classes in the trained BIOS model.

4.4 Quantifying Contrastive Decision-making

The *contrastive projection* $C(\vec{h}_{\mathbf{x}})_{y',y^*}$ (§3.3) derives a representation of the instance which only contains the information that differentiates two given classes. In other words, this is an amnesic intervention (at the last layer of the model's reasoning) which *removes* the information that cannot differentiate the two classes.

By treating the projection as an intervention, we can probe for the role that the contrastive decision-making has in the real process of the last layer in

[9] We use INLP (Ravfogel et al., 2020) for the amnesic operation, therefore the information we remove is linear, and we cannot guarantee the removal of non-linear information. As such, when we observe no change in the behavior, we can only say that the final layer made no use of the inspected concept, but it may have been used in a previous layer.

Fact (prediction)	Foil (**label**)	Highlight
attorney	no explicit foil	He has been involved in land transport for the past 13 years and has worked on various transport projects in Malta. He was appointed as chief officer for land transport within the Authority for Transport in Malta in 2010 where he was responsible for the regulation of driver training, testing and licensing, vehicle registration, goods transport, and passenger transport. In 2015 he moved back to the private sector and took on the role of General Manager of the local bus company, Malta Public Transport with his main responsibility being to oversee the transformation of the public transport service.
	accountant	He has been involved in land transport for the past 13 years and has worked on various transport projects in Malta. He was appointed as chief officer for land transport within the Authority for Transport in Malta in 2010 where he was responsible for the regulation of driver training, testing and licensing, vehicle registration, goods transport, and passenger transport. In 2015 he moved back to the private sector and took on the role of General Manager of the local bus company, Malta Public Transport with his main responsibility being to oversee the transformation of the public transport service.
physician	no explicit foil	She is a top medicine student whose academic achievements are sweet fruit of her labor. All seemed well until she reached her junior internship year, when one of her patients died under her watch. She was publicly humiliated in the aftermath. Her closest friends and family tried to lift her spirits up, but to no avail. She thought she was a failure. All she felt was the immense pressure boiling inside of her, and she can no longer contain it. Thus, on one fateful night, on the rooftop of her apartment, she decides to end her misery by taking her own life.
	surgeon	She is a top medicine student whose academic achievements are sweet fruit of her labor. All seemed well until she reached her junior internship year, when one of her patients died under her watch. She was publicly humiliated in the aftermath. Her closest friends and family tried to lift her spirits up, but to no avail. She thought she was a failure. All she felt was the immense pressure boiling inside of her, and she can no longer contain it. Thus, on one fateful night, on the rooftop of her apartment, she decides to end her misery by taking her own life.

Table 4: The top-1 results of ranking *highlights* contrastively given a particular foil, compared to doing so generally (i.e., where the foil is all other classes together) for examples where the model made a mistake. We consider the space of highlights to be all unigrams and bigrams, and rank the space by the change in behavior (via difference of normalized logits) following a masking intervention on the highlight.

the model. As before, we measure the change in behavior following the intervention. In this case, we measure global behavior change as symmetrized Kullback-Leibler divergence (D_{KL}).[10] Since the intervention already precisely maintains contrastive behavior, D_{KL} measures change in behavior in contrastive space.

Results for BIOS are in Table 5. We show the most affecting and least affecting foils for the same set of facts as in Table 2. A highly impacting foil in this table means that when predicting the fact, the differentiation of the fact from the foil *does not* significantly impact the decision (since removing the non-contrastive information greatly affects the decision), and vice-versa for the least impacting foils. For example, *attorney* and *paralegal* are semantically similar classes. Therefore, the differentiation between these classes is not a significant part of the decision making process for *attorney* predictions (2.195). Similar trends hold for *rapper* and *dj* (2.892), and opposite trends hold for semantically distinct classes, such as *attorney* and *dj* (0.528).

5 Case Study II: Analyzing MultiNLI

We apply a similar process to the *natural language inference* (NLI) task (RTE, Dagan et al., 2005) as another application of our methodology. The NLI task involves classifying a pair of sentences, a *premise* and a *hypothesis*, into whether the premise *entails*, *contradicts* or is *neutral* to the hypothesis. We analyze a model based on RoBERTa-Large which was fine-tuned on the MultiNLI (Williams et al., 2018) training set.

5.1 Evaluation via Data Staining

We are interested in assessing whether our method is sound and correct. We do so following a control setup, which allows us to perform a simple stress-test that verifies the method behaves as intended. We propose one method of doing so here.

Following Sippy et al. (2020), we design sanity-check evaluation of our methodology based on *data staining*. The process involves "staining" the training data with an artifact which is verifiably useful to solve the task, and then attempting to recover this artifact via the analysis. The stain may be any feature easily recoverable from the data instance, such as a specific word.[11]

It is possible to use data staining to assess contrastive explanation by introducing an artifact to the data which is only *contrastively* useful to the model: the "stain" is some artifact useful to differentiate between specific classes. The model is therefore encouraged to exploit the artifact in its decision making.

For our stain, we append each hypothesis in the MultiNLI training set with a prefix (the prefix being the stain), and this prefix is entirely correlated with a particular class—the *stained class*—but uncorre-

[10] Symmetrized D_{KL} referring to $D_{\text{KL}}(p \parallel q) + D_{\text{KL}}(q \parallel p)$.

[11] Previous work introduced the stain by altering the label distribution. This is marginally different from our case, where we introduce the stain by manipulating the input text.

Fact (y^*)	Top 3		Bottom 3	
	Foil	D_{KL}	Foil	D_{KL}
physician	surgeon	3.867	paralegal	0.847
	professor	2.400	dj	0.849
	nurse	2.027	photographer	0.875
attorney	professor	2.581	dj	0.528
	paralegal	2.195	personal trainer	0.576
	journalist	1.420	chiropractor	0.586
nurse	professor	2.386	dj	0.740
	physician	2.305	software engineer	0.747
	psychologist	1.662	rapper	0.763
yoga teacher	teacher	4.500	rapper	0.877
	nurse	2.480	surgeon	1.115
	psychologist	2.264	pastor	1.131
paralegal	attorney	4.680	surgeon	0.779
	accountant	2.295	professor	0.889
	interior designer	1.978	physician	0.993
rapper	dj	2.892	dietitian	0.725
	poet	2.705	yoga teacher	0.942
	comedian	1.931	architect	0.964
interior designer	architect	3.540	composer	0.869
	photographer	2.145	chiropractor	1.021
	journalist	2.054	pastor	1.150

Table 5: Results on the measurement of contrastive power in the decision process of the model's last layer for making *fact* predictions. This answers how much, on average, the model relies on differentiating *physician* from *surgeon* when making *physician* predictions. The measurement is *inverse* to the degree that the differentiation is dominating the decision process.

lated between the other two classes. This means that the stain is highly useful to the model to distinguish one class from the others: the stain is *contrastively useful* for or against the stained class, only.

We repeat our experiment three times, where each of the three classes (*entailment*, *contradiction* or *neutral*) stands for the stained class. Table 6 describes the stains used. The choice of prefix is intended to be semantically insignificant to the original text, though the stained task should be considered synthetic by its nature.

We analyze a `RoBERTa-Large` model fine-tuned[12] on the stained MultiNLI training set and attempt to recover the stains on the MultiNLI dev-set using our methodology, on the two axes of *foil ranking* and *factor ranking*. In all three cases, the stained models achieve high predictive test-set performance on the stained MultiNLI (above 97% accuracy). This high performance is expected, and indicates that the models indeed exploit the stains.

Stained Class	Label Prefix		
	entailment	contradiction	neutral
entailment	Indeed,	Though,	Though,
contradiction	Indeed,	No,	Indeed,
neutral	And,	And,	Though,

Table 6: Specification of the three data staining procedures. We prefix each hypothesis with a code word associated with each class. When we stain the entailment class, for example, all entailment examples are prefixed with "Indeed," while all other examples are prefixed with "Though," and thus the model is encouraged to use the stain *contrastively* only for or against the stained class.

Stained Class	Fact	Foil		
		entailment	contradiction	neutral
entailment	contradiction	**.0202**	—	.0037
	neutral	**.1012**	.0017	—
contradiction	entailment	—	**.0172**	.0059
	neutral	.0029	**.0668**	—
neutral	entailment	—	.0065	**.1058**
	contradiction	.0074	—	**.0791**

Table 7: Data staining results for foil ranking. The stained class emerges as the contrastive class for both possible facts, indicating that the model is indeed using the stain (artifact) primarily in contrast to the stained class.

Foil ranking. In this setting, we expect the stained class to be the most behaviorally influenced (in the contrastive space) by masking the stain (hypothesis prefix). We report that the experiments indeed achieved this result for all three variants (Table 7).

Highlight ranking. In this setting, we define our candidate factor space to be all tokens in the hypothesis, and we expect the first token (the stain) to be the most contrastively important evidence when either the fact or the foil is the stained class; otherwise, any other token should take its place. We measure this in accuracy, where a 'correct' hit denotes that the stain is highlighted when the fact or foil is the stained class, or that the stain is not highlighted when the stained class is not the fact nor the foil.

We perform the experiment for a random sample of 1000 MultiNLI instances. The results are 98.45% for the *entailment*-stained model, 96.9% for *contradiction*, and 96.1% for *neutral*. Examples of the experiment are in Table 8.

Note that the model is *not* guaranteed to make optimal use of the stain in every case—while we

[12] 10% of the instances that the model was trained on had their stain masked, in order to encourage such examples to be in-distribution.

encourage the model to do so by providing a useful stain, we cannot enforce the behavior. We determine that the model does not make optimal use of the stain by the fact that it does not reach optimal predictive performance on the stained class. Therefore, high but sub-optimal accuracy performance is within expectations.

5.2 Contrastive Power of NLI Concepts

Conceptual features have been discussed at length in the context of NLI tasks primarily with *erroneous heuristical behavior* as a result of spurious data correlations and artifacts (Naik et al., 2018; McCoy et al., 2019). For example, instances with negation words in the hypothesis are correlated with the *contradiction* label, and instances with high lexical overlap between premise and hypothesis with *entailment*. For example, models trained on these datasets tend to predict *contradiction* when words such as 'not', 'no', and so on are present in the hypothesis.

Literature in this area highlights considerable evidence that SNLI and MultiNLI models heavily leverage these concepts in their decision making process. But do they do so *contrastively*?

Refer to Table 9 for metrics on how these concepts are distributed in the training data of the model.

Overlap. Instances with the overlap concept are instances where all of the words in the hypothesis also exist in the premise (in any order), sans of stopwords. We remove the model's ability to represent the concept via amnesic probing, and probe for any change in behavior.

Previous work (Gururangan et al., 2018) have derived that the overlap concept is highly relevant in the model's reasoning process. Results in Table 10 shows that the *overlap* concept is overwhelmingly contrastive against the *neutral* class in particular. This aligns with Table 9, which shows that the concept is highly correlated with *entailment* and against *neutral*.

Hypothesis. The 'hypothesis' concept is an aggregation of all concepts which exist in the hypothesis, *without* the premise, and which are useful to the model for making predictions (Gururangan et al., 2018). For example, the length of the hypothesis, or the existence of particular expressions (e.g., negation, adjectives). Previous research (Poliak et al., 2018) demonstrates that this concept is often useful to NLI models. We recover this concept by training a hypothesis-only baseline on MultiNLI, and assigning the concept to instances where the baseline was correct. As before, we analyze this concept's contrastive role via amnesic probing.

Results in Table 10 show that this concept is *not* strongly contrastive to either foil in particular. This aligns with how this concept is distributed in the data (Table 9), since it is equally correlated against *neutral* and *contradiction*.

However, of note is that while the role of this concept is not strong in MultiNLI decisions on average, it is strong on SNLI decisions, implying that the model is more likely to exploit features that are only encoded in the hypothesis (as opposed to including premise reasoning) on SNLI data than on MultiNLI. This aligns with previous research that SNLI is prone to these issues (Gururangan et al., 2018).

Negation. The negation concept is represented by examples that have a negation word in the hypothesis.[13] This concept is highly correlated with the *contradiction* label in the MultiNLI dataset, and previous work has shown this concept to be represented in NLI models' reasoning process (Gururangan et al., 2018). Otherwise, the concept is similarly correlated with both entailment and contradiction.

For this concept, we test for contrastive behavior via manually annotated counterfactuals, in place of amnesic probing, as (1) it is easier to do so, in comparison to the previous concepts; and (2) amnesic probing uncovered no signal for this concept, such that either the concept is truly not relevant for the model, or the role of the concept in the model's reasoning exists prior to the intervention (since it is at the last layer of the model, and not the input layer). As previous work showed that this concept is important, we suspect that the the latter is true.

Given an example which the model predicted as *entailment* or *neutral*, which does not contain a negation, we manually paraphrased the hypothesis to include a negation. We collected 90 examples for each of *entailment* and *neutral* in this way.[14] We then proceed to probe the model for behavioral changes between the instances before–and–after the intervention, treating the negated example as the context for the fact and the original example as

[13] Such as: not, nobody, never, nothing, no, none, neither, or nowhere.

[14] Made available in our repository.

Stained Class	Prediction (Fact)	Foil	Text & Highlight	Passes?
entailment	contradiction		P: Ramses II did not build it from stone but had it hewn into the cliffs of the Nile valley at a spot that stands only 7 km (4 miles) from the Sudan border, in the ancient land of Nubia.	
		entailment	H: *Though,* Ramses II ordered that it be made out of stone and not hewn into the cliffs.	✓
		neutral	H: *Though,* Ramses II ordered that it be made out of stone and not hewn into the cliffs.	✓
entailment	neutral		P: yeah that's the World League	
		entailment	H: *Though,* that's the World League that you can join.	✓
		contradiction	H: *Though,* that's the World League that you can join.	✓
entailment	entailment		P: Their ideas and initiatives can be implemented at the local and national levels.	
		contradiction	H: *Indeed,* locally and nationally, their ideas can be applied.	✓
		neutral	H: *Indeed,* locally and nationally, their ideas can be applied.	✓

Table 8: Examples of the data staining experiment design. Each hypothesis is augmented with a prefix (*italics*) conditioned on the stained class; *entailment* with 'Indeed,' and the others with 'Though,'. For each foil, we test whether the stain is highlighted. If the stained class is the fact or the foil, we hypothesize that the stain *should* be highlighted, thus it "passes" the test.

	In data	entailment	contradiction	neutral
Overlap	6.07%	65.6%	28.7%	5.8%
Hypothesis	53.2%	49.7%	22.5%	27.8%
Negation	15.05%	18.9%	63.2%	17.9%

Table 9: Distribution of examples with the concept for each *predicted* label (i.e., across model predictions) via the training set. For example, 65.6% of examples with the overlap concept were predicted as *entailment* by our model. 'In data' refers to the percentage of examples with the concept in the data.

Concept	Fact		Foil	
		entailment	contradiction	neutral
Overlap	entailment	—	0.006	0.433
Hypothesis (MultiNLI)	entailment	—	−0.005	−0.031
Hypothesis (SNLI)	entailment	—	0.505	0.463
Negation	contradiction	0.195	—	0.051

Table 10: Results for the degree of contrastive role that NLI concepts have in the MultiNLI model's decision process on the MultiNLI dev set (and on SNLI for the 'Hypothesis' concept), by measuring change in behavior following intervention.

the context for the foil.

Indeed, the negation caused an increase in assigned probabilities to *contradiction*, in line with previous work. Results in Table 10 imply that the model utilizes the negation concept as evidence for *contradiction* rather than *entailment*, on average, more than *neutral*. We leave additional research into why this may be the case to future work.

Summary. We analyze the contrastive role of three well-researched concepts in NLI. While previously shown as useful to NLI models, we now understand whether they are useful against a particular foil. Two concepts do have a prominent contrastive role, while one does *not*, showing that some concepts are indeed useful to the model without clear contrast.

5.3 Highlight Ranking

As before we rank highlights in a given highlight space—in this case, single tokens in the hypothesis—by their contrastive relevance to the decision, for each foil.

Qualitative results are in Table 11 where we report the top-1 highlight for each foil. We focus on examples where the model predicted the wrong label, in the interest of understanding the error in reasoning. We see a trend in which the one of the two possible class foils "dominates" the highlight explanation in the general case (where the foil is both classes together). *Therefore the contrastive perspective reveals additional detail in the model reasoning process, which is otherwise missing.*

In example (4) we see a case where the dominating foil was *not* the gold label, and thus, explaining contrastively reveals new insight on why the model specifically preferred its predicted class *over the gold label*. The implication of this result, we theorize, is that the model could not derive that soccer games are played on fields, since this common knowledge is missing from the premise—and the location (a bar) is additional information, making the example *neutral* rather than *entailment*. However, the model recognizes that both instances involve a soccer game, therefore it is *neutral* and not *contradiction*.

6 Related Work

Our approach for deriving contrastive explanations relies on an intervention in the model's working mechanism, either via manipulating the input or the representation of the model. This interventional approach follows several recent works in the field of NLP (Giulianelli et al., 2018; Meyes et al., 2020; Vig et al., 2020; Elazar et al., 2020; Feder et al.,

Id	Fact (prediction)	Foil (**label**)	Text & Highlight
(1)	entailment		P: A nun uses her camera to take a photo of an interesting site.
		no explicit foil	H: A nun taking photos of a interesting site outside.
		contradiction	H: A nun taking photos of a interesting site outside.
		neutral	H: A nun taking photos of a interesting site outside.
(2)	neutral		P: A couple bows their head as a man in a decorative robe reads from a scroll in Asia with a black late model station wagon in the background.
		no explicit foil	H: A light black late model station wagon is in the background.
		entailment	H: A light black late model station wagon is in the background.
		contradiction	H: A light black late model station wagon is in the background.
(3)	neutral		P: Girl plays with colorful letters on the floor.
		no explicit foil	H: The girl is having fun learning her letters.
		entailment	H: The girl is having fun learning her letters.
		contradiction	H: The girl is having fun learning her letters.
(4)	neutral		P: Three men with blue jerseys try to score a goal in soccer against the other team in white jerseys and their goalie in green.
		no explicit foil	H: Some men with jerseys are in a bar, watching a soccer match.
		entailment	H: Some men with jerseys are in a bar, watching a soccer match.
		contradiction	H: Some men with jerseys are in a bar, watching a soccer match.

Table 11: Several examples of the *highlight ranking* procedure from the e-SNLI dev-set. These examples illustrate the difference between highlights that are explicitly contrastive to another class, and those which are not—which are generally dominated by one of the contrastive perspectives, and do not show the full picture. We consider only single-token highlights here, but the method is applicable for any highlight space.

2020), and is justified by accumulating empirical evidence for the inability to draw causal interpretation from statistical associations alone (Hewitt and Liang, 2019; Tamkin et al., 2020; Ravichander et al., 2020; Elazar et al., 2020).

Kaushik et al. (2020); Gardner et al. (2020) generate contrast sets that differ in selected aspects from naturally-occurring sentences (e.g. by sentiment or syntactic properties). The manual generation allows for greater expressivity, on the expense of longer and more labor-intensive process. Other recent works try to automate the generation process, such as by conditioned language generation (Wu et al., 2021). Note that these works are ultimately intervening on the gold labels, and not on model predictions, making them less suitable for contrastive analysis.

Rathi (2019) propose a model-agnostic contrastive explanation scheme based on Shapley values. They offer a local explanation, unlike our global method. In addition, our approach employs behavioral interventions, while Rathi (2019) do not. Others have raised concerns regarding feature importance methods based on Shapley-values (Kumar et al., 2020); the implicit foil of such methods can be unintuitive to human explainees.

In computer-vision, many have studied the generation of counterfactual explanations or counterfactual data points. Hendricks et al. (2018) proposed a method that provides natural language counterfactual explanation of image classification decisions. They have relied on a model that proposes potential counterfactual evidence, followed by a verifier that is based on human-provided image description. As their method relies on pre-existing explanation model and human descriptions, there is no guarantee the explanation it provides are related to the model's reasoning process. Sharmanska et al. (2020) used GANs to generate examples representing minority groups, to improve fairness measures. This work, like other works in vision, relies on the continuous input, which is not present in natural-language applications.

The area of contrastive explanations in ML and NLP is relatively new. Recently, Jacovi and Goldberg (2020a) offer a method of deriving highlights which contain what portion of the input which ultimately flips the model decision from *foil* to *fact* (the final model prediction); Ross et al. (2020) offer a method of generating minimally edited counterfactuals that similarly flip the model decision. Such methods can be understood as particular types of interventions on the highlighted or edited portions of the input, and are therefore orthogonal to our work, which proposes a contrastive framework to understand results from intervention experiments.

7 Discussion and Future Work

Given the relative novelty of contrastive explanations in NLP and ML research, there remain several open questions in this topic, which are potential future directions of research.

Evaluating explanations. Evaluation of explanations in neural NLP models is a difficult and ongoing research area. While the mathematical interpretation of our methodology is sound, some questions on evaluation remain. How can we choose foils such that explanations are more comprehensible to humans? Similarly, what should causal factors satisfy, in terms of their *social attribution* (Jacovi and Goldberg, 2020a)? Our work provides sanity-check evaluations based on data staining. Additionally, our experimental results align with prior intuition on well-researched models and tasks.

Types of interventions. Our amnesic interventions, e.g., masking highlights, and latent space concept removal, generate counterfactual representation for behavioral analysis. However, amnesic counterfactuals do not provide a complete picture of how humans comprehend decision processes: the information communicated in the explanation is contextual in the type of intervention applied. Currently, literature is limited for the trade-offs between various types of interventions (and types of counterfactuals) in explanation. Additional research in this area may answer what exactly do amnesic counterfactuals do and do not communicate, and how to categorize classes of possible counterfactuals by their utility.

8 Conclusion

We proposed a framework for producing contrastive explanations in text classification. Our framework is based on selection from a pre-defined space of candidate factors and model decisions. We presented a behavioral analysis where results were measured as change in model behavior following an intervention. We propose a theoretical (mathematical) modification of model behavior into a contrastive reasoning process, which allows the behavioral change to be interpreted in the context of a particular alternative decision. Our framework is flexible and generalizable, as is demonstrated via our experiments with two text classification tasks on BIOS and MultiNLI. Overall, we observe, via quantitative and qualitative evaluations, that a contrastive context to explanations makes them easier to understand, and allows finer-grained communication of aspects of model behavior.

References

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Maria De-Arteaga, Alexey Romanov, Hanna M. Wallach, Jennifer T. Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Cem Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 120–128. ACM.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2020. Amnesic probing: Behavioral explanation with amnesic counterfactuals.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2020. CausaLM: Causal model explanation through counterfactual language models.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 107–112. Association for Computational Linguistics.

Xiaochuang Han, Byron C. Wallace, and Yulia Tsvetkov. 2020. Explaining black box predictions and unveiling data artifacts through influence functions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5553–5563, Online. Association for Computational Linguistics.

Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Generating counterfactual explanations with natural language. *CoRR*, abs/1806.09809.

Germund Hesslow. 1988. The problem of causal selection. In Denis J. Hilton, editor, *Contemporary Science and Natural Explanation: Commonsense Conceptions of Causality*. New York University Press.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 2733–2743. Association for Computational Linguistics.

Denis J Hilton. 1988. Logic and causal attribution. In *Part of this chapter is the text of a paper read at the symposium," Attitudes and Attribution: A Symposium in Honour of Jos Jaspars," convened at the Annual Conference of the British Psychological Society, Sheffield, England, Apr 3-5, 1986.* New York University Press.

Denis J Hilton and Ben R Slugoski. 1986. Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological review*, 93(1):75.

Alon Jacovi and Yoav Goldberg. 2020a. Aligning faithful interpretations with their social attribution.

Alon Jacovi and Yoav Goldberg. 2020b. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. 2020. Learning the difference that makes A difference with counterfactually-augmented data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2017. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav).

Siwon Kim, Jihun Yi, Eunji Kim, and Sungroh Yoon. 2020. Interpretation of nlp models through input marginalization.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR.

I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle A. Friedler. 2020. Problems with shapley-value-based explanations as feature importance measures. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 5491–5500. PMLR.

Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2016. Rationalizing neural predictions. *CoRR*, abs/1606.04155.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Richard Meyes, Constantin Waubert de Puiseau, Andres Posada-Moreno, and Tobias Meisen. 2020. Under the hood of neural networks: Characterizing learned representations by functional neuron populations and network ablations. *CoRR*, abs/2004.01254.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38.

Tim Miller. 2020. Contrastive explanation: A structural-model approach.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Shubham Rathi. 2019. Generating counterfactual and contrastive explanations using SHAP. *CoRR*, abs/1906.09293.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard H. Hovy. 2020. Probing the probing paradigm: Does probing accuracy entail task relevance? *CoRR*, abs/2005.00719.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA. ACM.

Alexey Romanov, Maria De-Arteaga, Hanna M. Wallach, Jennifer T. Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Cem Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Kalai. 2019. What's in a name? reducing bias in bios without access to protected attributes. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4187–4195. Association for Computational Linguistics.

Alexis Ross, Ana Marasović, and Matthew E. Peters. 2020. Explaining nlp models via minimal contrastive editing (MiCE).

Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2662–2670.

Viktoriia Sharmanska, Lisa Anne Hendricks, Trevor Darrell, and Novi Quadrianto. 2020. Contrastive examples for addressing the tyranny of the majority. *CoRR*, abs/2004.06524.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Jacob Sippy, Gagan Bansal, and Daniel S Weld. 2020. Data staining: A method for comparing faithfulness of explainers. In *Proc. of ICML Workshop on Human Interpretability in Machine Learning (WHI)*.

Alex Tamkin, Trisha Singh, Davide Giovanardi, and Noah D. Goodman. 2020. Investigating transferability in pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 1393–1401. Association for Computational Linguistics.

Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 10–19. ACM.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. 2020. Causal mediation analysis for interpreting neural NLP: the case of gender bias. *CoRR*, abs/2004.12265.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 11–20. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S. Weld. 2021. Polyjuice: Automated, general-purpose counterfactual generation.

Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel, and Max Welling. 2017. Visualizing deep neural network decisions: Prediction difference analysis. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.