NEGATION IS NOT ENOUGH: A DEEPER ANALYSIS INTO PREMISE NEGATION ARTIFACTS IN SNLI

A PREPRINT

Tanner O'Rourke Department of Natural Sciences
University of Texas, Austin
two297@my.utexas.edu

December 6, 2025

ABSTRACT

While state-of-the-art Natural language inference (NLI) models are able to capture detailed variance in Question-Answering (QA) settings and achieve high accuracies, they still carry tendency to over-generalize hypothesis-only cues, simple negation, and lexical overlap. Simply put, while a premise-hypothesis pair of Prem: *Prem: "there is an image here" / Hyp: "No image can be seen"* is an easily predicted contradiction, a premise-hypothesis pair of Prem: *Prem: "there is not an image here" / Hyp: "No image can be seen"* is both an underrepresented and biased example case due to simple placement of negation. In fine-tuning a baseline ELECTRA-small model Clark et al. [2020] on the stanford NLI dataset, we find this baseline model performs expectedly better on high lexical overlap and root object part of speech negation, we find that SNLI not only contains scarce *premise-negation* examples across all parts of speech, which the model performs poorly outright on, but also that the model overgeneralizes a "hypothesis-negation $\rightarrow$ contradiction", most visible when hypothesis-negation examples are labeled as non-contradictory. Driven by these findings, we first perform (i) part-of-speech (PoS) granularity analysis of negation, (ii) hypothesis vs. premise negation bias reduction Bowman et al. [2015], and (iii) design a two-stage intervention with a goal of increasing accuracy in negation scopes while maintaining overall accuracy metrics *without model augmentation* (with simple models such as ELECTRA-small, small additions such as residual correction head or hyperparameter tweaking). We then build out an augmentation pipeline that (1) injects targeted contrast examples into both training and evaluation splits to populate underrepresented slices Gardner et al. [2020a], (2) reruns slice analyses to define a simple error-based reweighting scheme Bhagavatula et al. [2020], and (3) retrains with both augmentation and slice-aware weighting integrated, along with analysis of baseline, augmented, and augmented + reweighted models on both i.i.d. SNLI and our contrast sets. We find with this approach that improving coverage of rare negation patterns keeps overall model metrics stable, raises performance on certain PoS patterns within and between premises and hypotheses, yet premise-only negation remains difficult to improve.

1 Introduction

Natural language inference (NLI) benchmarks such as SNLI play a central role in developing sentence-level reasoning systems, yet are well known to contain lexical artifacts and annotation shortcuts that undermine an ability to evaluate true semantic competence. Prior work has shown that models can rely on easily exploitable features—negation words, bag-of-words similarity, and frequent antonyms—to achieve deceptively high accuracy Gururangan et al. [2018]. Gardner et al. [2020b] advocates for small contrast sets, human-crafted dataset perturbations, as a way to probe model's local decision boundary and reveal brittleness hidden by i.i.d. accuracy.

This work adopts this perspective by to a standard ELECTRA-small model fine-tuned on SNLI. NPAS serves. We fine-tune an ELECTRA-small discriminator on SNLI using the provided HuggingFace pipeline and then study its behavior through lexically defined slices and contrast sets. Concretely, we analyze slices based on: (i) **NPAS (Negation-Perturbation-Analysis with Slices** (i.e, where negation appears) as a diagnostic lens through which we identify

the model's specific weaknesses and the dataset gaps along multiple axes, (ii) lexical overlap between premise and hypothesis, and (iii) distribution analysis of labels vs. negation constructs between the dataset and model predictions Our baseline analysis reveals that hypothesis-level negation and high lexical overlap are handled considerably well, yet the premise negation in SNL produces sharp accuracy drops. Additionally, we find biased artifacts that promote overgeneralization of contradiction when faced with a negating hypothesis. These patterns mirror known issues in NLI models: underrepresented linguistic structures and a reliance on shallow lexical heuristics rather than compositional reasoning.

Building on these diagnostics, we explore a two-stage, slice-guided augmentation. First, we construct targeted contrast examples and inject them into the training and evaluation sets, with templates specifically designed to (a) populate rare slices like premise negation and (b) break straightforward lexical heuristics (e.g., hypothesis-negation equals contradiction, negated-yet-entailed hypotheses). Second, after re-running the slice analysis on the augmented model, we derive a simple error-based weighting scheme that up weights high-error patterns during a second round of training, yielding a final "augmented+reweighted" model. This setup allows us to ask not only which heuristics the model actually uses, but also how far we can push robustness with low-cost, slice-informed modifications to the training distribution. Our goal was to not only to measure robustness but also to demonstrate and analyze how lightweight, interpretable interventions can actively reshape a model's learned decision surfaces.

2 Expiremental Setup

Our experiments follows a three-stage pipeline utilizing the Stanford Natural Language Inference (SNLI) dataset as our base corpus and keeping a static model architecture fixed to ELECTRA-small, avoiding any form of model augmentation (e.g., residual correction heads, additional heads, or architectural modifications).

2.1 Model and Training Configuration

Our training procedure implemented the ELECTRA-small discriminator architecture through HuggingFace Transformers. The model consists of a 12-layer encoder with 256 hidden units, 4 attention heads, and a classification head producing logits for entailment, neutral, contradiction.

Training Procedure

- Optimizer: AdamW
- Learning rate: default
- Batch size: 38
- Epochs: 6 (all training and augmentation phases)
- Max sequence length: 128 (all training and augmentation phases)
- Max training examples: 100,000 (used to gather clean measurements while still utilizing high example volume)
- Loss: Cross-entropy, optionally modified by slice-specific weights during the second training phase
- Hardware: Home PC (RTX 3070, Ryzen 5700X, 32gb RAM)

Injected and Reweighted Models

- Augmented model: trained on SNLI + injected contrast examples.
- Augmented+weighted model: trained with the same data plus a loss reweighting factor applied to slices with high baseline error (e.g., premise-negation examples, lexical-overlap contradictions, age-incompatible events).

2.2 Deriving NPAS Part-of-Speech Negations

A central goal of our analysis is to understand where negation operates within the syntactic structure of SNLI examples and how a fine-tuned model responds to negation in different grammatical positions. Negation does not behave uniformly across sentences: a negated subject ("no child...") introduces different semantic meaning than a negated predicate ("did not laugh...") or a negated quantifier ("not many people..."). Our work hypothesizes that the *location of negation within the dependency tree* is essential for interpreting artifacts.

- Explicit negation: we detect explicit lexical negators (commonly used in negation detection, such as "not", "never", "no"), as well as their contracted versions ("don't", "can't", etc)

- Dependency Relation: `spaCy` annotates negation scopers using the dependency relation label `neg`. We utilize this to find the syntactic head that is being negated to uncover the "negation anchor" within a structured span of text

By combining these concepts, we can assign negation instances to a *syntactic region* that reflects the grammatical role of the negation token. This creates the following dependency categories:

- "root": negation modifies the main predicate (e.g., "did not run")
- "subject": negation modifies the subject noun phrase (e.g., "no boy laughed")
- "object": negation attaches to thedirect or indirect object (e.g., "ate no dinner")
- "locative": negation appears inside an 'oblique' phrase (e.g., "not in the room")
- "attribute": negation appears in the copular or adjective structure (e.g., "is not happy")
- "quantifier": negation modifies the determiner or quantifier (e.g., "not many people")

Our analysis derives utilizes a derivation of each usage of negation seperately, as well as negation in the hypothesis or premise i.i.d. Additionally, we derive tangential dimensions of negation in the SNLI data, outlined in our analysis.

3 Analysis of Lexical Heuristics

Premise-Negation sparcity. Our analysis is primarily driven by the lack of premise negation examples. As shown in Table 1, there are considerably less of these examples. *SNLI premises typically come from image captions* (Flickr30k), which usually describe *present, observable facts*. The captions rarely include logical operators, non-literal language, or negation cues. Thus, we posit-ed that these sets are not entirely representative of natural language premises.

pattern	# exs (/ 100,000)	accuracy (%)
Overall	10,000	0.854
hypothesis (no negation)	9792	85.29%
hypothesis (negation)	9943	85.44%
premise (negation)	208	89.42%
premise (no negation)	57	75.44%
Low overlap	5,635	85.0%
Medium overlap	3,976	85.2%
High overlap	389	92.0%

Table 1: Accuracy and number of examples by negation pattern, and overlap score (equal to the Jaccard Score between premise and hypothesis

Lexical overlap. We also studied the premise of lexical overlap, where we saw the model is slightly stronger on high-overlap examples (0.92 vs. ≈ 0.85 elsewhere), but there is no catastrophic failure; performance is relatively stable across lexical overlap. Among the high-overlap examples that also contain explicit negation cues in the hypothesis, the slice is extremely small ($n = 2$) and both examples are classified correctly. This suggests vanilla lexical overlap and hypothesis negation alone aren't where the model fails most dramatically.

Frequency and distribution of negation. To reason more precisely about negation, we augmented each example with binary indicators for whether the premise and hypothesis contains an explicit negation cue:

- NONE: no explicit negation in either sentence;
- HYP_ONLY: negation in the hypothesis only;
- PREM_ONLY: negation in the premise only;
- BOTH: negation in both premise and hypothesis.

This allows us to summarize how often each pattern appears and the empirical label distributions in gold and predictions.

Pattern	n	Gold label dist.			Pred label dist.		
		ent	neu	con	ent	neu	con
NONE	9,738	0.337	0.327	0.335	0.343	0.333	0.324
HYP_ONLY	205	0.117	0.137	0.746	0.117	0.122	0.761
PREM_ONLY	54	0.333	0.370	0.296	0.333	0.278	0.389
BOTH	3	0.333	0.000	0.667	0.000	0.333	0.667

Table 2: Baseline label distributions by negation pattern on SNLI validation (n = 10,000). "Gold" and "Pred" are empirical distributions over labels within each pattern.

Several facts are immediately visible:

1. **Negation is extremely rare in SNLI, especially in premises.** The NONE pattern covers 97.4% of the validation set. Only 2.1% of examples have negation in the hypothesis (HYP_ONLY + BOTH), and only 0.6% have negation in the premise (PREM_ONLY + BOTH).
2. **Hypothesis-level negation is strongly associated with contradiction in the dataset, and the model mirrors this.** In the HYP_ONLY slice, 74.6% of gold labels are CONTRADICTION, and the model predicts CONTRADICTION 76.1% of the time. This suggests that the model has largely learned the dataset prior "negation in the hypothesis $\rightarrow$ likely contradiction" rather than inventing such a pattern on its own.
3. **Premise-level negation has a much more mixed label distribution in the data, but the model pushes it toward contradiction.** In PREM_ONLY, the gold labels are relatively balanced (33.3% entailment, 37.0% neutral, 29.6% contradiction), whereas the model's predictions skew towards contradiction (38.9%) and away from neutral (27.8%). Given how few PREM_ONLY examples exist, this misalignment suggests the model is overusing a shortcut "any negation equals contradiction" in precisely the region where the dataset does *not* support that prior.

Negation Reliance Index for hypothesis vs. premise. To quantify to what extent the model overuses negation as a cue for specific labels, we devised a Negation Reliance Index (NRI). For a given binary negation flag $z \in \{\text{neg}, \text{no-neg}\}$ and target label ℓ, define

$$\Delta_{\text{gold}}(\ell) = P_{\text{gold}}(\ell \mid z = \text{neg}) - P_{\text{gold}}(\ell \mid z = \text{no-neg}),$$
$$\Delta_{\text{pred}}(\ell) = P_{\text{pred}}(\ell \mid z = \text{neg}) - P_{\text{pred}}(\ell \mid z = \text{no-neg}),$$
$$\text{NRI}(\ell) = \Delta_{\text{pred}}(\ell) - \Delta_{\text{gold}}(\ell).$$

Intuitively, the NRI tells us if the model's dependence of label on negation is stronger than what the dataset itself exhibits. Concretely $\text{NRI}(\ell) > 0$ means the model is dependent of label ℓ on negation by ℓ percentage points more than what the dataset itself exhibits.

The below Table reports NRI values for the target label CONTRADICTION when negation is flagged in the hypothesis vs. premise. The probabilities are computed by aggregating the pattern-level distributions in baseline negation patterns table into two groups (neg vs. no-neg) for each side.

Flag	$P_{\text{gold}}(\text{CON} \mid \text{neg})$	$P_{\text{gold}}(\text{CON} \mid \text{no-neg})$	$P_{\text{pred}}(\text{CON} \mid \text{neg})$	$\text{NRI}(\text{CON})$
Hypothesis negation	0.745	0.335	0.760	0.025
Premise negation	0.316	0.343	0.404	0.099

Table 3: Negation Reliance Index (NRI) for CONTRADICTION. Rows show how much the presence of negation in the hypothesis or premise changes the probability of predicting CONTRADICTION, relative to the change observed in the gold labels.

For **hypothesis negation**, the dataset strongly correlates negation with CONTRADICTION ($\Delta_{\text{gold}} \approx 0.41$), yet the model only *slightly* exaggerates this (NRI ≈ 0.025). In other words, the model mostly mirrors the dataset prior. However for**premise negation**, the dataset does *not* makes contradiction more likely. If anything, $P_{\text{gold}}(\text{CON} \mid \text{prem\_neg} = 1)$ is slightly *lower* than $P_{\text{gold}}(\text{CON} \mid \text{prem\_neg} = 0)$. In contrast, the model's predicted contradiction probability jumps from roughly 0.333 to 0.404 when premise negation is present, leading to a substantially positive NRI ≈ 0.10. This indicates that the model is injecting an additional "premise negation $\rightarrow$ contradiction" bias that is not warranted by the dataset.

Error patterns motivating augmentation. Qualitative inspection of misclassified examples in the premise-negation slices supports this picture. In PREM_ONLY examples where the gold label is NEUTRAL, the baseline often predicts CONTRADICTION, even when the hypothesis simply adds compatible information. For instance:

> *Premise*: There is not an image here.
> *Hypothesis*: There is only a single color displayed instead of an image.
> *Gold*: NEUTRAL *Predicted*: CONTRADICTION (baseline)

Despite the hypothesis being a plausible elaboration of the premise, the presence of a negation cue in the premise appears to encourage the model to default to contradiction.

Taken together, the baseline analysis leads to the following narrative:

1. On the majority of the data (NONE pattern), the model is well-calibrated and does not overuse contradiction.
2. In hypothesis-negation cases, the model largely reflects a real dataset bias: negation in the hypothesis frequently corresponds to contradiction in SNLI, and the model learns this mapping.
3. In premise-negation cases, however, the dataset provides *few* examples, with a balanced gold distribution that often favors NEUTRAL or ENTAILMENT, while the model exhibits a stronger-than-justified preference for CONTRADICTION.

This combination of scarcity of premise-negation examples, misalignment between gold and predicted label distributions in PREM_ONLY, and elevated NRI for CONTRADICTION under premise negation directly motivates the augmentation strategy in the following section. We specifically target regions with premise-level negation—especially cases where the gold label is neutral or entailment—both by adding challenge examples and by reweighting these slices during training, with the goal of softening the spurious prior that "any premise-level negation implies contradiction."

4 Data Augmentation for Premise-Level Negation

The baseline analysis showed that SNLI contains very few examples with premise-level negation, and that the model tends to over-predict CONTRADICTION for these cases, especially when the gold label is NEUTRAL. To directly target this artifact, we applied two complementary interventions: (i) targeted contrastive data augmentation, and (ii) loss reweighting for premise-negation slices.

Contrastive challenge examples. We first constructed a small contrast set focused on premise-level negation (grew to $\approx 50 \times 3$ examples, this was a core point of exploration). Starting from SNLI training examples with explicit negation in the premise, we manually edited or generated additional hypotheses that (i) showcased negation in cases with underrepresentation in the dataset, (ii) preserved the core event described by the premise, and (iii) produced non-CONTRADICTION labels where appropriate (e.g., adding negated hypotheses that are entailed by a negated premise, or neutral elaborations that remain compatible with it). We added these contrastive examples to **both** the training and validation data on a $70/30$ split after a deterministic shuffle, increasing the effective coverage of PREM_ONLY and BOTH patterns and, in particular, adding examples where premise-level negation should map to ENTAILMENT or NEUTRAL rather than CONTRADICTION.

Slice-aware loss reweighting. After validating negation distributions through a dry training run, we applied margin-dependent reweighting schema in the training loss. Examples with premise-level negation (either PREM_ONLY or BOTH)—and especially those tagged with premise-side negation on the root, subject, or quantifier—were upweighted relative to the majority NONE slice. PoS regions which performed *better* than overall accuracy were not upweighted. The goal was to encourage the model to devote more capacity to these underrepresented configurations and makes errors on premise-negation examples more costly during learning. The resulting model was then trained on the same base architecture, optimizer, and hyperparameters as the baseline, but on an augmented dataset with modified per-example weights.

Overall, the augmentation approach was explicitly designed to (a) expose the model to more varied premise-negation semantics, and (b) counteract the spurious heuristic that "any premise-level negation implies contradiction" that we identified in baseline inference.

4.1 Experiments and Results

On the SNLI validation set (9,888 + 122 = 10,000 injected examples after filtering unlabeled items), the augmented model achieved an overall accuracy of 86.22%, slightly higher than the baseline 0.8538. Table **??** compares key negation slices before and after augmentation.

Pattern	Model	n	Accuracy
Overall	Baseline	10,000	0.8538
	Augmented	9,888	0.8622
NONE	Baseline	9,791	0.8530
	Augmented	9,590	0.8636
HYP_ONLY	Baseline	170	0.9235
	Augmented	210	0.8667
PREM_ONLY	Baseline	37	0.7568
	Augmented	75	0.6667
BOTH	Baseline	2	0.5000
	Augmented	13	0.8462

Table 4: baseline vs. augmented model accuracy by negation pattern on SNLI validation

The first noticeable trend is that the model remained strong (improved) on the corpus as a whole, with a modest accuracy gain from 85.30% to 86.36%. This was an unexpected behavior that we hypothesize is correlated with the addition of greater variation in the examples, and indicates that the augmentation does not degrade performance in the majority regime. Secondly, the coverage of **rare negation configurations improved**: the BOTH slice increases from only 2 to 13 examples, and its accuracy rose from an inadequate 50.0% to 85.0%. However, accuracy on PREM_ONLY drops from 0.76 to 0.67 when evaluated on the new, larger set of premise-negation-only examples, suggesting that the augmented model still finds these examples difficult.

The label distributions by negation pattern for the augmented model (shown below in table) show that the core artifact identified in the baseline persists in a refined form. The dataset strongly favors CONTRADICTION (71.4%) for HYP_ONLY, and the augmented model predicts CONTRADICTION 77.6% of the time. In the NONE case, gold and predicted distributions remain closely matched. In PREM_ONLY, however, the gold labels are mixed — 22.7% ENTAILMENT, 42.7% NEUTRAL, 34.7% CONTRADICTION — while the model pushes a larger fraction into CONTRADICTION (52.0%) and underpredicts NEUTRAL (21.3%). This is also corroborated by the Negation Reliance Index (NRI) values

Pattern	n	Gold label dist.			Pred label dist.		
		ent	neu	con	ent	neu	con
NONE	9,590	0.343	0.333	0.325	0.341	0.335	0.324
HYP_ONLY	210	0.143	0.143	0.714	0.124	0.100	0.776
PREM_ONLY	75	0.227	0.427	0.347	0.267	0.213	0.520
BOTH	13	0.846	0.000	0.154	0.769	0.077	0.154

Table 5: Label distributions by negation pattern for the augmented model (SNLI validation).

For CONTRADICTION conditioned on hypothesis negation, the augmented model exhibits only a mild over-reliance (NRI ≈ 0.057), broadly, which tracks the dataset's strong coupling between hypothesis negation and contradiction. In contrast, for premise negation the NRI for CONTRADICTION increases to about 0.147: the gold contradiction rate slightly *decreases* when premise negation is present, but the model's predicted contradiction rate increases from roughly 0.33 to 0.47. Confusion matrix analysis showed that around half of gold-NEUTRAL examples with premise negation are predicted as CONTRADICTION, while most gold-ENTAILMENT and gold-CONTRADICTION examples are correctly classified.

In a deeper PoS-based negation analysis, slices reveal that the augmented model is performative when negation appears in the hypothesis across syntactic regions (accuracies typically $\approx 0.84 - 0.92$), but remains inaccurate when negation appears in the premise root or subject (accuracies around $\approx 0.58 - 0.59$) and in premise quantifier negations (0.40 on a very small slice). This suggests that while our augmentations improved overall performance and coverage of

rare patterns, it did not fully resolve the deeper issue: the model continues to over-interpret premise-level negation as evidence for contradiction, especially when that negation affects the core proposition.

In conclusion, the augmented model continued to exhibit a pronounced bias toward predicting CONTRADICTION when the premise contains negation, however performed slightly stronger in aggregate metrics and better supported rare negation regimes. This artifact is precisely the behavior our analysis set out to debunk, and highlights that correcting such biases may require more targeted or structural interventions than simple contrastive augmentation and slice-aware reweighting.

4.2 Limitations of Slice-Guided Augmentation

Our analysis and mitigation strategy had several limitations when compared to an operational approach. First, all results are tied to the SNLI dataset, which contains extremely few premise-level negation examples even after augmentation. Our manually constructed contrast sets increased coverage, but remain small and targeted; they were designed to "stress-test" specific failure modes rather than approximate the full distribution. As a result, we level with the fact that these findings may not directly transfer to other NLI datasets or more diverse domains.

Second, our negation detection and localization pipeline relies on spaCy parsing and a set of hand-designed rules to place negation into coarse syntactic regions. Parsing errors or mismatches between dependency labels and semantic scope took up most of the development time and inevitably introduced noise into the slice definitions. While the noise is partly mitigated due to a dependency-span-focused approach to aggregating negation and trends, it must be true that various examples are likely mis-categorized, especially in more complex sentences.

Third, our intervention was intentionally focused on dataset-intervention and slice-based loss reweighting. Prior work has shown that more sophisticated model-level techniques—such as ensemble-based debiasing, training separate "artifact expert" models, or adding residual correction heads that explicitly model the difference between a full model and an artifact-only baseline—can yield larger gains on challenge sets and better robustness overall. We expect that such architectures enhancements would more positively improve accuracy on negation slices, however our goal in this project work to freeze model-level changes and expose dataset artifacts on a pretrained model, rather than maximize performance, and demonstrate how lightweight augmentations interact with those artifacts.

References

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. *cs.CL*, 2020. URL https://arxiv.org/abs/2003.10555.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi:10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075/.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna 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. Evaluating NLP models via contrast sets. *CoRR*, abs/2004.02709, 2020a. URL https://arxiv.org/abs/2004.02709.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. Abductive commonsense reasoning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Byg1v1HKDB.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. *CoRR*, abs/1803.02324, 2018. URL http://arxiv.org/abs/1803.02324.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna 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. Evaluating NLP models via contrast sets. *CoRR*, abs/2004.02709, 2020b. URL https://arxiv.org/abs/2004.02709.