Measuring and Improving Attentiveness to Partial Inputs with Counterfactuals

Yanai Elazar[1,2] **Bhargavi Paranjape**[2]* **Hao Peng**[3]* **Sarah Wiegreffe**[1,2]*
Khyathi Raghavi Chandu[1] **Vivek Srikumar**[4] **Sameer Singh**[5] **Noah A. Smith**[1,2]
[1]Allen Institute for AI
[2]University of Washington
[3]University of Illinois Urbana-Champaign
[4]University of Utah
[5]University of California, Irvine
yanaiela@gmail.com

Abstract

The inevitable appearance of spurious correlations in training datasets hurts the generalization of NLP models on unseen data. Previous work has found that datasets with paired inputs are prone to correlations between a specific part of the input (e.g., the hypothesis in NLI) and the label; consequently, models trained only on those outperform chance. Are these correlations picked up by models trained on the full input data? To address this question, we propose a new evaluation method, Counterfactual Attentiveness Test (CAT 🐱). CAT uses counterfactuals by replacing part of the input with its counterpart from a different example (subject to some restrictions), expecting an *attentive* model to change its prediction. Using CAT, we systematically investigate established supervised and in-context learning models on ten datasets spanning four tasks: natural language inference, reading comprehension, paraphrase detection, and visual & language reasoning. CAT reveals that reliance on such correlations is mainly data-dependent. Surprisingly, we find that GPT3 becomes *less* attentive with an increased number of demonstrations, while its accuracy on the test data improves. Our results demonstrate that augmenting training or demonstration data with counterfactuals is effective in improving models' attentiveness. We show that models' attentiveness measured by CAT reveals different conclusions from solely measuring correlations in data.[1]

1 Introduction

Reliance on spurious correlations compromises NLP models' abilities to generalize across different domains and tasks (Naik et al., 2018; McCoy et al., 2019). Intuitively, spurious correlations are features that are useful in the training data but are unreliable in general (Eisenstein, 2022). Different approaches have been proposed for measuring model reliance on spurious correlations at various granularities, e.g., the appearance of certain tokens in a text, the length of a text, etc. (Sinha et al., 2021; Gardner et al., 2021).

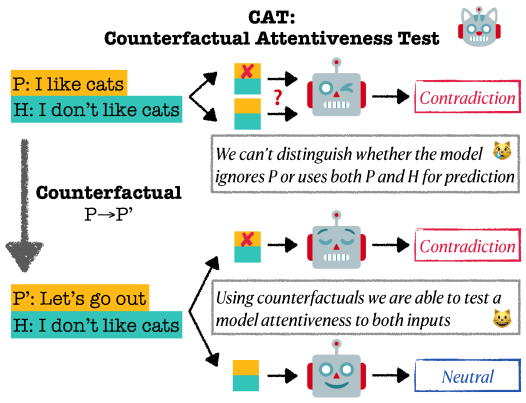

Figure 1: We are interested in quantifying model's attentiveness to part of the input. We propose CAT, an evaluation that replaces the premise P with a counterfactual P'. A change in behavior to the counterfactual input indicates the model is attentive to the premise, otherwise, the model relies solely on the hypothesis to make a prediction.

Spurious correlations in datasets are very common. For instance, the *hypothesis-only baseline* (Poliak et al., 2018; Gururangan et al., 2018) was designed to measure spurious correlations in natural language inference (NLI) datasets by training a model only on one part of the input (e.g., the hypothesis), effectively discarding information deemed critical for performing the task. Surprisingly, it performs significantly better than random guessing, despite never seeing the premise. These results suggest that the hypothesis-only baseline has taken the shortcuts between the hypotheses and the labels during training. However, they should not be interpreted as evidence that models trained on the full inputs (*full-input models*) also take these shortcuts. The question that then arises is *how to*

*Equal contribution.
[1]Code available at: https://github.com/yanaiela/partial_input.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 3603–3623
November 12-16, 2024

quantify the reliance of full-input models on the spurious correlations present in the hypotheses.

We propose a method for measuring models' reliance on partial input: COUNTERFACTUAL ATTENTIVENESS TEST, (CAT, §3). We present an overview of the approach in Figure 1. CAT is simple, intuitive, and is based on the idea of counterfactuals from the causal inference literature (Pearl, 2009). Importantly, it only requires a labeled dataset for the task and black box model predictions. To test a model that performs a tasks with paired-inputs (tasks that consist of at least two components, e.g., NLI), we query the model, and replace one part of the input (e.g., the premise) with that part from another instance randomly drawn from the same data split. Intuitively, this perturbation results in the *default* label for that task (typically a label that signifies no relation between the two inputs, e.g. *neutral* in NLI). As such, we focus on the subset of instances where the model predicts *non-neutral*.[2] If the prediction changes, it suggests that the model is *attentive* to the perturbed input. Conversely, a model that keeps the prediction unchanged is likely to be *not attentive*, as it insensitive to the counterfactual. Following this intuition, our metric measures *attentiveness* by calculating the percentage of predictions that *changed* from the non-neutral labels on the original instances.

We conduct extensive experiments on four different tasks, ten datasets, and 15 different models, using two setups: supervised and in-context learning (§4). We first extend previous work to obtain new results using the *partial input baseline*—a model trained only on partial inputs, that quantifies spurious correlations in the data. Then, using our attentiveness metric, we show that even in datasets with spurious correlations between parts of the inputs and the labels, models do not always rely on them (§5). Finally, we study whether data augmentation using our counterfactuals improves models' attentiveness and find that is often the case (§6). Our results indicate that the appearance of — often unavoidable — spurious correlations in the data do not indicate models rely on them, and propose a simple and easy-to-use test for measuring it.

2 Background & Related Work

In this section we review the *hypothesis-only* baseline in NLI and its extensions to other tasks: the *partial input baseline*. We then discuss a related work by Srikanth and Rudinger (2022) and highlight the similarities and differences to our work.

Partial Input Correlations The hypothesis-only baseline refers to a supervised classifier trained only on the hypotheses in an NLI dataset, removing the premises from the original paired inputs (Gururangan et al., 2018; Poliak et al., 2018; Tsuchiya, 2018). Predicting an NLI label on a single-sentence input is ill-defined, but nonetheless, classifiers trained solely on the hypotheses on datasets such as SNLI (Bowman et al., 2015) are able to generalize and achieve better-than-random accuracy on the corresponding test sets. This result indicates that the dataset used to train the hypothesis-only classifier contains predictive information about the labels. This phenomenon is often referred to as *spurious correlations* because, in principle, both the premise and hypothesis should be required to infer the label. For instance, Poliak et al. (2018) found that words such as 'no', 'sleeping', and 'cat' were often used for generating hypotheses that contradict the premises, while 'instrument', and 'touching' were used more with entailed hypotheses. Other works discovered similar behavior in other tasks and datasets (Kaushik and Lipton, 2018a; KV and Mittal, 2020; Trivedi et al., 2020; Mihaylov et al., 2018; Hessel and Lee, 2020).

NLI Models Reliance on the Hypothesis Recently, Srikanth and Rudinger (2022) raised the question of whether models trained on the *full input* data instances are in fact using such correlations (e.g., in NLI, ignoring the premise when making a prediction). They proposed to use counterfactuals to study whether full-input models rely on such heuristics. Importantly, Srikanth and Rudinger (2022) manually constructed the counterfactuals, which limits the scale of such analysis. Their study uses RoBERTa-base and two NLI datasets (SNLI (Bowman et al., 2015) and δ-NLI (Rudinger et al., 2020)) which led to the conclusion that models are *not* relying on such heuristics. Our study covers a larger scope, including more tasks, datasets and setups (we also consider in-context learning), and reaches a more nuanced conclusion. We find that models trained on full inputs may still rely on such heuristic, while it is task and data-dependent. We provide a detailed comparison with Srikanth and Rudinger (2022), and summarize the similarities and differences in Appendix E.

[2] Other datasets and tasks may have different *neutral* labels. We provide more details on these labels in §3.

Counterfactuals for Evaluating NLP Models Counterfactuals were studied before in NLP. Some works use counterfactuals to study models' robustness to counterfactual perturbations (Glockner et al., 2018; Kaushik et al., 2020; Dev et al., 2020; Gardner et al., 2020; Wu et al., 2021; Fryer et al., 2022; Teney et al., 2020). However, these works typically construct the counterfactuals manually, that involves expensive manual labor, in contrast to our method that produces counterfactuals automatically. Finally, there is a line of research that modifies neural networks' parameters or intermediate states, to study high-level concepts, which are harder to create counterfactuals for, in the input-level (Serrano and Smith, 2019; Elazar et al., 2021; Feder et al., 2022; Wu et al., 2022; Geiger et al., 2021, 2022).

3 Quantifying Attentiveness

We seek a method that measures model attentiveness to a part of the input in paired-input tasks, such as the premise or hypothesis in NLI. Formally, assume an input x contains two parts: $x : (x_1, x_2)$, and let $\mathcal{M}_{x \to y}$ be a model trained to predict y from x. We are interested in measuring the causal effect of x_1 on model $\mathcal{M}$. If such effect is prominent, we conclude that the model relies on x_1 for inference,[3] otherwise it ignores it. By measuring such effect on a dataset we quantify the extent by which the model relies on x_1. Cases where such effect is large are especially interesting when the underlying training data contains correlations between x_2 and y, since the model does not rely on such correlations during training. We describe the *partial input baseline* and a naive approach for measuring attentiveness and explain why they do not fit as such metric. We then describe our method, CAT, in §3.3. We summarize these approaches in Table 1.

3.1 Partial Input Baseline

As discussed in §2, previous approaches such as the hypothesis-only baseline (Poliak et al., 2018) train a model solely on part of the input, x_2 (e.g., the hypotheses in NLI and paragraphs in RC, $\mathcal{M}_{x_2 \to y}(x_2)$). However, this approach does not tell us whether a *full-input* model ($\mathcal{M}_{x \to y}(x)$) also relies on parts of the inputs as it is a different model; it only tells us whether parts of the inputs contain predictive information about the label. Therefore, the partial input baseline solely discovers spurious correlations in a dataset, and is not suitable for studying full-input models' behavior.

3.2 Naive Solution

Instead of relying on a partial input model, we first introduce another naive approach to quantify a model's reliance on x_2, and explain why it is not valid for answering our research question. Naively, one may use a full-input model $\mathcal{M}_{x \to y}(x)$ and evaluate the model on a partial input $\mathcal{M}_{x \to y}(x_2)$. Then, we would test whether the predictions are equal, which would indicate that the model only relies on x_2. However, attributing attentiveness (or inattentiveness) is problematic since we cannot disentangle the cause of such behavior, which can result from either the model's attentiveness, or from the model's behavior on out-of-domain data distribution (Fong and Vedaldi, 2017; Hooker et al., 2019).

3.3 Counterfactual Attentiveness Test

We propose an alternative solution that can be computed automatically, and provides a better estimate of the model's reliance on parts of the input. We call this method COUNTERFACTUAL ATTENTIVENESS TEST (CAT 🤖). The method is inspired by *counterfactuals* from causal inference to ask "what would happen if part of the input was different?" (Pearl, 2009). Instead of completely removing a part of the input, e.g., x_1, we replace it with x_1' and evaluate the model's prediction on $\mathcal{M}_{x \to y}(x_1', x_2)$. Intuitively, replacing x_1 by x_1' is likely to result in a non-related pair. As such, in almost all cases[4] the new label is *neutral* in the case of NLI.[5]

As we are interested in measuring models' attentiveness with a change in prediction, we cannot estimate it in cases where the model predicts neutral on the original (x_1, x_2) pair. We thus only consider cases where the prediction on the original pair is non-neutral. Since the predictions on the original (x_1, x_2) pairs may differ between models, we cannot compare them directly. However, in Appendix §D, we compute the attentiveness on the subset where models' predictions are identical and show that since the original predictions are often similar, the trends are consistent.

Attentiveness Metric We use accuracy to measure the attentiveness of a model on the counter-

[3] We leave it to future work to explore problems where the input has even more structure (i.e., more than two parts), but conceptually the generalization is straightforward.

[4] We confirm this assumption empirically in §4.2.

[5] Each task and dataset have their own corresponding labels. We discuss those labels in §4.

Method	Train	Eval.	Eval. Ex
Full input	$\mathcal{M}_{x\to y}(x)$	$\mathcal{M}_{x\to y}(x)$	P: The other men shuffled. H: The other men were shuffled around.
Prior work (§3.1)	$\mathcal{M}_{x_2\to y}(x_2)$	$\mathcal{M}_{x_2\to y}(x_2)$	H: The other men were shuffled around.
Naive partial-inputs (§3.2)	$\mathcal{M}_{x\to y}(x)$	$\mathcal{M}_{x\to y}(x_2)$	H: The other men were shuffled around.
COUNTERFACTUAL ATTENTIVENESS TEST (§3.3)	$\mathcal{M}_{x\to y}(x)$	$\mathcal{M}_{x\to y}(x'_1, x_2)$	P: The dog barked. H: The other men were shuffled around.
Counterfactual attentiveness data augmentation (§6)	$\mathcal{M}_{(x_1,x_2)\to y}(x)$	$\mathcal{M}_{(x_1,x_2)\to y}(x)$	P: The other men shuffled. H: The other men were shuffled around.
	$\mathcal{M}_{(x'_1,x_2)\to y}(x)$	$\mathcal{M}_{(x'_1,x_2)\to y}(x)$	P: The dog barked. H: The other men were shuffled around.

Table 1: Comparing the methods we consider. Unlike prior work ("hypothesis-only baselines"), our proposed counterfactual method does not change the model. We additionally show with counterfactual training that training on examples of counterfactual partial-inputs can improve performance on them. $\mathcal{M}$ refers to a model, (x) is a paired input that consist of the two variables (x_1, x_2), similarly to $x'_2 : (x'_1, x'_2)$, and y is the model's prediction.

factual instances. Specifically, we consider as an initial set all of the instances (x_1, x_2) from the development set. We obtain the model's prediction and only keep the instances where the model predicts a non-*neutral* label. Then, we generate a counterfactual by randomly sampling an instance (x'_1, x'_2) from the same data split, and combining x_2 with x'_1.[6] Finally we compute how often the model changes its prediction from the initial pair on this subset. A higher score indicates the model is more attentive to the full input, which signifies the model does not rely on partial inputs for making predictions. For each instance (x_1, x_2) in the dataset, we sample $k = 5$ counterfactuals x'_1 from the data and report the mean and standard deviation accuracy across the samples.

4 Experimental Setup

We consider four diverse tasks and ten English datasets. We also consider two learning strategies: supervised learning through fine-tuning using 12 models from six model families, and in-context learning (ICL) using nine models from three model families. Table 3 summarizes all configurations.

4.1 Tasks and Datasets

We use the standard train-development splits[7] for each dataset. In the ICL setup, we randomly sample k instances from the training sets.

Natural Language Inference We consider three datasets for NLI: MNLI (Williams et al., 2018), WANLI (Liu et al., 2022), and RTE (Dagan et al., 2005). MNLI and WANLI contain three labels: *entailment*, *neutral*, and *contradiction*, while RTE only contains *entailment* and *non-entailment*.

Paraphrase Detection We consider two datasets for paraphrase detection (PD): Quora Question Pairs (QQP; Sharma et al., 2019) and PAWS (Zhang et al., 2019). These datasets contain questions that are labeled as *paraphrase* or *non-paraphrase*.

Reading Comprehension We consider three datasets for reading comprehension (RC) that include a subset of questions that are not answerable i.e., contain a *no-answer* label: SQuAD2.0 (Rajpurkar et al., 2018), DuoRC (Saha et al., 2018), and NewsQA (Trischler et al., 2017).

Visual & Language Reasoning We consider two datasets for visual and language reasoning tasks (VLR). VQA2.0 (Goyal et al., 2019) contains questions and images with multiple choice answers. NLVR2 (Suhr et al., 2019) consists of two images and a statement where the task is to predict whether the statement is true for the images.

4.2 Counterfactual Default Labels

CAT generates counterfactuals by automatically sampling two inputs from different instances. This makes an assumption that no relation would hold between the inputs in the new counterfactual instance (as shown in the examples in Table 2), together with the respective label. Such relations are different between tasks and datasets, which we now specify: *neutral* in 3-class NLI (e.g. MNLI), *non-entailment* in 2-class NLI (e.g. RTE), *not-paraphrase* in the PD, *no-answer* in RC, *false* in NLVR2, and *no-answer* in VQA.

Label-Flip Assumption Verification We verify CAT's assumptions by hand-annotating 50 counterfactually-paired instances per dataset for each of our four tasks. We annotate whether a representative model's predicted label for an instance is incorrect for the counterfactual instance and

[6] As opposed to previous work (Srikanth and Rudinger, 2022) that relies on hand-crafted counterfactuals to probe the model, we rely on much easier-to-collect counterfactuals, which can be randomly sampled from the data and do not require manual curation.

[7] Except for WANLI which does not have a development set, for which we randomly sample 5K instances from the training set and use them for evaluation.

Task	Dataset	Counterfactual Example			Acc
		x_1	x_2	y	
NLI	MNLI	Perhaps North Africans and eastern Europeans peopled the Ligurian coast, while the Adriatic and south may have been settled by people from the Balkans and Asia Minor.	You can find more information from the senior executive's plan.	neutral	100
	WANLI	The best known is the building that houses the National Gallery in Trafalgar Square, London, designed by Sir Charles Barry and completed in 1843.	A great many of the villages in the area are in a state of repair, and some of them are actually inhabited.	neutral	100
	RTE	Schroder Investment Management has indicated its intention to accept Revival's offer to buy retailer Marks & Spencer.	There are 32 pandas in the wild in China.	non-entailment	100
PD	QQP	How do you delete messages on Snapchat?	Which are some of the best companies to work for?	non-paraphrase	100
	PAWS	The Arieşul Mare River is a tributary of the Vâlcea River in Romania.	Also steam can be used and need not be pumped.	non-paraphrase	100
RC	SQuAD 2.0	Hypersensitivity is an immune response that damages the body's own tissues. They are divided into four classes (Type I – IV) based on the mechanisms involved... These reactions are mediated by T cells, monocytes, and macrophages.	What tracts does commensal flora help pathogens thrive in?	no answer	98
	DuoRC	South Boston teenager Jason Tripitikas is a fan of martial arts films and awakens from a dream of a battle between the Monkey King and celestial soldiers in the clouds. He visits a pawn shop in Chinatown to buy Wuxia DVDs and discovers a golden staff. On his way home, Tripitikas	What is Ana's profession?	no answer	100
	NewsQA	Iran's parliament speaker has criticized U.S. President-elect Barack Obama for saying that Iran's development of a nuclear weapon is unacceptable. Iranian President Mahmoud Ahmadinejad has outlined where he thinks U.S. policy needs to change. Ali Larijani said Saturday that Obama should apply his campaign message of change to U.S. dealings with Iran.	What does the U.N drug chief advocate?	no answer	98
VLR	VQA 2.0		What color are this person's shoes?	no answer	92
	NLVR2		There is a smartphone in the right image.	false	96

Table 2: An example of a counterfactual pair (x_1', x_2) from each of the datasets we consider. The label y is assigned automatically to the counterfactual data. "Acc." indicates the % of instances from our manual annotation experiment (§4.2) where the assumption holds that the model's predicted label should change from the original to the counterfactual instance if the model is attentive.

Setup	Task	Dataset	Model
Supervised	NLI PD RC	RTE, MNLI, WANLI QQP, PAWS SQuAD2.0, DuoRC, NewsQA	BERT, RoBERTa, DeBERTa, T5, T5-v1.1, Flan-T5
	VLR	VQA2.0, NLVR2	BLIP, ViLT
ICL	NLI PD	RTE, MNLI, WANLI QQP, PAWS	GPT-3, T5, Flan-T5

Table 3: Details about the configurations we use. We experiment with ten datasets and 11 model families.

should thus change if the model is attentive. Overall we find the assumption to be true most of the time: 100% of instances in MNLI, WANLI, RTE, QQP, PAWS and DuoRC; 98% in SQuAD2.0 and NewsQA, 96% in NLVR2, and 92% in VQA 2.0.[8] We provide some randomly-selected instances from each dataset in Table 2, along with the accuracy from our manual evaluation.

4.3 Models

Supervised Learning We experiment with six text-only models and two multimodal models which were trained on some pretraining corpora, and then fine-tuned on a respective supervised dataset (e.g., MNLI). We use three masked language models: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), DeBERTa v3 (He et al., 2021). and three T5 models (Raffel et al., 2020), with both the v1.0, and v1.1 versions (v1.0 was first pre-trained, and then fine-tuned on supervised training data from GLUE (Wang et al., 2018), and QA tasks, while v1.1 was only pre-trained as a language model). Additionally, we also use Flan-T5 (Chung et al., 2022), T5 models that were fine-tuned on 1,836 tasks with instructions. We use the base and large versions of each model. For multimodal models, we experiment with BLIP (Li et al., 2022), and ViLT (Kim et al., 2021).

In-Context Learning We experiment with three types of model families that were shown to perform well in zero-shot and few-shot settings using in-context learning (ICL). The first, text-davinci-003, is a GPT-3 model that was first trained on code and then was additionally trained to follow instructions (Ouyang et al., 2022). We refer to this model simply as GPT-3. Since the model is a commer-

[8] Cases where the assumption fails for VQA are primarily yes/no questions where the answer is still no after the image is shuffled (4/25). When the questions are shuffled, the assumption holds 100% of the time (25/25).

cial product, we do not have full information of its training data, and other crucial reproducibility details. For instance, it is possible that the evaluation data we test the model on were seen by the model. However, it is reasonable to assume that the model has not seen the paired input from CAT, as these are randomly constructed pairs, and those results can be trusted. In addition, we experiment with two open-sourced models: T5 (Raffel et al., 2020), and Flan-T5 (Chung et al., 2022). Flan-T5 was initialized from the T5 model, and finetuned on 1,836 supervised tasks accompanied by instructions.

4.4 Learning Strategies

Supervised Learning through Finetuning We consider the fine-tuning setup to train a model per task and dataset. Each model in these experiments is trained for 10 epochs, following Bandel et al. (2022) to reduce reliance on lexical overlap heuristics. We use the same learning rate for all models.[9]

In-Context Learning We follow the prompt format in Raffel et al. (2020) for the T5 models. For GPT-3 and Flan-T5, we prepend instructions to each instance (described in Table 14 in the Appendix).[10] We consider both zero-shot and few-shot settings. Each demonstration includes examples from all labels, and we consider $k \in \{0, 1, 2, 3\}$ number of tuples, which translates into $\{0, 2, 4, 6\}$ or $\{0, 3, 6, 9\}$ depending on the number of labels in a dataset. We use greedy decoding for generating the answers for each of these models.

5 Results

Standard Evaluation We begin by training models on their corresponding datasets. For example, DeBERTa-large achieves 87.73%, 91.28%, and 76.72% accuracy on RTE, MNLI, and WANLI, respectively.[11] The full results are in Appendix A.

Partial Input Baseline We experiment with the partial input baseline (§3.1) and train with only parts of the inputs. We extend previous work (Gururangan et al., 2018; Kaushik and Lipton, 2018b) and perform extensive experiments with different types of models, tasks, and datasets. Every experiment corresponds to the evaluation performance of a model trained on the respective training set from the same data. The results are in Appendix B.

	NLI		PD	RC
Dataset	**RTE**	**MNLI**	**PAWS**	**SQuAD2.0**
BERT-base	$99.8_{\pm 0.3}$	$74.6_{\pm 0.3}$	$99.3_{\pm 0.2}$	$98.0_{\pm 0.0}$
BERT-large	$100.0_{\pm 0.0}$	$64.3_{\pm 0.1}$	$85.9_{\pm 0.3}$	$98.0_{\pm 0.2}$
RoBERTa-base	$99.8_{\pm 0.3}$	$73.4_{\pm 0.5}$	$80.5_{\pm 0.6}$	$98.3_{\pm 0.1}$
RoBERTa-large	$100.0_{\pm 0.0}$	$69.8_{\pm 0.2}$	$78.7_{\pm 0.3}$	$98.2_{\pm 0.2}$
DeBERTa-base	$97.0_{\pm 1.6}$	$78.8_{\pm 0.4}$	$99.9_{\pm 0.1}$	$97.6_{\pm 0.1}$
DeBERTa-large	$100.0_{\pm 0.0}$	$68.6_{\pm 0.3}$	$99.1_{\pm 0.1}$	$98.0_{\pm 0.2}$
T5-base	$99.4_{\pm 0.7}$	$81.3_{\pm 0.1}$	$99.0_{\pm 0.1}$	$99.1_{\pm 0.1}$
T5-large	$100.0_{\pm 0.0}$	$71.5_{\pm 0.4}$	$100.0_{\pm 0.0}$	$99.5_{\pm 0.1}$
T5-1.1-base	-	$46.9_{\pm 0.6}$	-	$99.4_{\pm 0.0}$
T5-1.1-large	-	$79.4_{\pm 0.3}$	-	$99.4_{\pm 0.1}$
FLAN-T5-base	$99.9_{\pm 0.3}$	$81.8_{\pm 0.3}$	$98.3_{\pm 0.2}$	$99.4_{\pm 0.1}$
FLAN-T5-large	$100.0_{\pm 0.0}$	$77.1_{\pm 0.2}$	$100.0_{\pm 0.1}$	$99.5_{\pm 0.1}$

Table 4: CAT results on natural language inference (NLI), paraphrase detection (PD) and reading comprehension (RC). Results are calculated by computing the mean attentiveness of five counterfactuals per instance, and their standard deviation. Higher numbers indicate better model attentiveness. We mark results for models whose standard performance is not significantly better than random with '-'.

Hypothesis-only MNLI models achieve as high as 61.66% accuracy with T5-1.1-large, a much higher accuracy than the majority baseline (35.4%). On the other hand, models trained on RTE and WANLI only achieve 10.5% and 7.4% points better than majority. For QA, the strong performance of the question-only models is particularly noteworthy. The best results with DeBERTa-large is 98.5% on SQuAD2.0, 73.2% on DuoRC, and 96.8% on NewsQA. These models are pre-trained on a large amount of text data in the same domain as the respective datasets and are likely to select the correct answer span from a random passage based on their parametric knowledge and grammar heuristics.

Our results support and extend prior findings by showing that for a diverse class of models and a variety of tasks, models trained on partial inputs are strong. Next, we revisit the idea that these results indicate standard models trained on such datasets also relies on such heuristics.

5.1 Counterfactual Attentiveness Test

Next, we report the results of our Counterfactual Attentiveness Test on the different setups. Overall, the main trend we observe in our experiments is that attentiveness is data and setup specific. The results are summarized in Table 4 for the supervised textual tasks, Table 5 for the visual reasoning tasks,

[9] Except for the smaller models (T5-small and base) as they were unable to learn using smaller learning rates.

[10] Preliminary experiments find that including instructions substantially improves both the accuracy and attentiveness.

[11] Our MNLI results are comparable to those by (He et al., 2021), but on RTE ours is worse by 5%. This is likely because we do not perform extensive hyper-parameter search. Note that we achieve a reasonable performance nonetheless.

Dataset	VQA v2	NLVR2
BLIP (image)	$65.1_{\pm 0.0}$	$95.9_{\pm 0.3}$
BLIP (text)	$89.6_{\pm 0.0}$	$96.0_{\pm 0.4}$
ViLT (image)	$63.6_{\pm 0.0}$	$93.3_{\pm 0.3}$
ViLT (text)	$89.3_{\pm 0.0}$	$93.1_{\pm 0.3}$

Table 5: CAT results on visual reasoning (VLR). Results are calculated by computing the mean attentiveness of five counterfactuals per instance, and their standard deviation. Higher numbers indicate better model attentiveness. The portion of the input that is perturbed is indicated in parentheses.

and Table 6 for ICL classification tasks.[12]

Supervised Models Attentiveness scores are consistent across models, e.g., on RTE all models are highly attentive with average scores ranging from 97.0% to 100%. On the other hand, models trained on MNLI are much less attentive, with scores ranging from 46.9% (T5-1.1-large) to 81.8% (FLAN-T5-base). In the case of VLR, we test the attentiveness of models to both parts of the input, once to the image and once to the text. The results indicate that models are more attentive to the images, than to the texts. For instance, in VQA v2, when perturbing the images, BLIP achieves 65.1%, but when perturbing the texts it gets 89.6%.

ICL We report the ICL attentiveness results in Table 6. The results in this setup follow a similar trend to the supervised models, where attentiveness results are data dependent. On RTE, models are highly attentive, performing 100% on CAT in most cases, and above 98% in the other cases. On MNLI however, the attentiveness varies from 28.3% (T5-3B, one-shot), 62.5% (GPT3, 3-shot) to 98.9% (Flan-T5-XL, 0-shot).

5.2 Partial Input Baseline as Indication of Attentiveness?

Next, we study the relationship between the *partial input* baseline performance as a method to estimate the spurious correlations in a dataset, and the attentiveness of a model and its ability to overcome these correlations. This relationship is presented in Figure 2. We plot the results of the supervised models, trained on the different considered datasets. On the x-axis, we plot the difference between the trained partial input baseline model and random

[12]The full results are presented in the Appendix (Table 15).

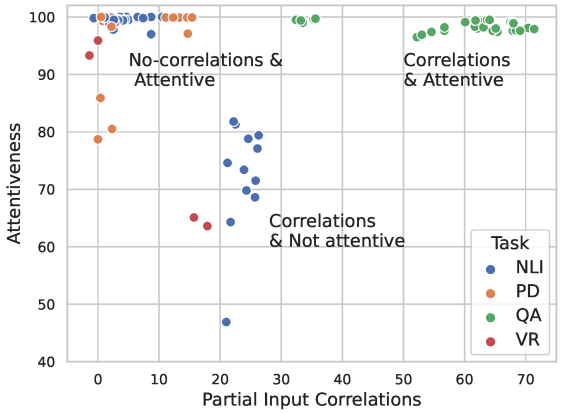

Figure 2: Attentiveness as a function of partial input correlations for the supervised models. Higher attentiveness values (y axis) indicate better attentiveness to counterfactual inputs. Higher values on the partial input correlations indicate correlations between a part of the input and the label. It is computed as the score on the *partial input* baseline, subtracting the majority score.

accuracy (*partial input correlations*), and on the y-axis, the model's attentiveness score.

First, we highlight the left part of the figure where the spurious correlations are small: models trained on datasets with less than a 10% gap with the partial input baseline over the majority label are mostly robust, with attentiveness scores as high as 100% (with 3 outliers, where the smallest one achieves 78.7%). This is unsurprising, as low partial input correlations indicate fewer spurious correlations in the dataset, so full-input models should be able to detect counterfactual inputs more readily. Next, we observe scattered patterns, typically clustered based on tasks and datasets, and overall worse attentiveness. For instance, the middle cluster consists of NLI models trained on MNLI, with high partial input correlations (between 20-30%) and low attentiveness (between 56.9-81.8%). Interestingly, models trained on RC datasets appear to be attentive (above 95%), while also having high partial input correlation scores (32.4%-71.4%). These results highlight two findings: (1) low partial input correlations in the training data result in high attentiveness, and (2) **high partial input correlations are not a good indicator of standard model behavior**: some models (trained on certain datasets) rely on such heuristics, whereas others do not.

5.3 ICL: The Role of Demonstrations

The standard accuracy (Table 9 in the Appendix) shows two trends: (1) more demonstrations im-

Model	#Tuples	NLI		PD
		RTE	MNLI	PAWS
GPT3	0	$99.5_{\pm 0.4}$	$92.9_{\pm 0.3}$	$99.9_{\pm 0.0}$
GPT3	1	$100.0_{\pm 0.0}$	$66.0_{\pm 0.3}$	$100.0_{\pm 0.0}$
GPT3	2	$100.0_{\pm 0.0}$	$62.5_{\pm 0.5}$	$100.0_{\pm 0.0}$
GPT3	3	$100.0_{\pm 0.0}$	$62.0_{\pm 0.3}$	$99.9_{\pm 0.0}$
Flan-T5-base	0	$100.0_{\pm 0.0}$	$66.5_{\pm 0.3}$	$99.9_{\pm 0.0}$
Flan-T5-base	1	$98.1_{\pm 1.4}$	$48.6_{\pm 0.4}$	$100.0_{\pm 0.0}$
Flan-T5-large	0	$100.0_{\pm 0.0}$	$94.8_{\pm 0.2}$	$100.0_{\pm 0.0}$
Flan-T5-large	1	$99.7_{\pm 0.3}$	$96.2_{\pm 0.3}$	$100.0_{\pm 0.0}$
Flan-T5-XL	0	$100.0_{\pm 0.0}$	$98.9_{\pm 0.1}$	$100.0_{\pm 0.0}$
Flan-T5-XL	1	$100.0_{\pm 0.0}$	$98.6_{\pm 0.2}$	$100.0_{\pm 0.0}$
T5-base	0	$100.0_{\pm 0.0}$	$76.6_{\pm 0.3}$	-
T5-base	1	-	$51.4_{\pm 3.4}$	-
T5-large	0	$100.0_{\pm 0.0}$	$72.9_{\pm 0.5}$	-
T5-large	1	$99.3_{\pm 0.6}$	$43.5_{\pm 5.5}$	-
T5-3B	0	$100.0_{\pm 0.0}$	$70.5_{\pm 0.3}$	-
T5-3B	1	-	$28.3_{\pm 4.5}$	-

Table 6: CAT performance using ICL. We report the mean and standard deviation across five counterfactuals per instance. Models whose standard performance is not significantly better than random are marked with '-'.

prove GPT's accuracy, but generally decrease the accuracy for others. Can we rely on these results as the main measure for model behaviors? ICL attentiveness results (Table 6) tell a different story. While on RTE, and the two PD datasets the attentiveness performance stays the same - near perfect, the attentiveness on MNLI and WANLI decreases when more demonstrations are used.

One hypothesis for GPT-3's low attentiveness is benchmark contamination (Rogers, 2023; Jacovi et al., 2023). If the model was exposed to the datasets we investigate and memorized their labels, this could result in high standard accuracy (which we observe). The low attentiveness we observe similarly supports the memorization hypothesis, since our perturbations are far less likely to have been observed by the model during training, so a model that has memorized rather than generalized would not do well on them. These results highlight the importance of reporting metrics such as attentiveness when clean evaluation sets are not available.

6 CAT Augmentation

Considering the simplicity and automatic nature of CAT, we reuse the same sampling process for data augmentation to improve non attentive models. We follow a similar procedure as in the evaluation to augment the training data: apart from the regular training set, for every x_1, x_2 pair, where the label y is different than the default label (e.g. *neutral* for NLI), we sample x_1' from the training data, assigned the default label.[13]

Setup	Model	Acc.	Attentiveness
Finetune	T5-large	89.3	$71.5_{\pm 0.4}$
	+augmentation	89.3	**99.7**$_{\pm 0.0}$
ICL	GPT3, 4 tuples, half neutral	78.1	$69.3_{\pm 0.2}$
	GPT3, 3 tuples	77.4	$66.2_{\pm 0.3}$
	+augmentation	66.4	**76.7**$_{\pm 0.3}$

Table 7: Standard accuracy performance on the respective development sets models are trained on, and the counterfactual accuracy of our metric.

The following experiments study the models that perform poorly on our metric in the standard setup. We focus on MNLI since models are the least attentive on this dataset, and experiment once with T5-large in the supervised setup, and with GPT-3 in the ICL setup, using the original data (or demonstrations), and add the counterfactual instances. The labels of counterfactuals match the default label - *neutral* for MNLI. We report both accuracy on the standard development set that verifies the augmentation does not degrade model performance and attentiveness. For each dataset, we report the performance of the vanilla model (those trained in §5 and used for CAT in §5.1) and the augmented model. The results are in Table 7, which can be directly compared to the results of training on the standard datasets in Tables 8 and 9 in the Appendix.

For supervised learning, augmentation has little impact on standard accuracy but improves attentiveness. T5-large achieves 71.5%, and 99.7 in attentiveness before and after the augmentation, respectively (and 89.3% accuracy both before and after augmentation on the MNLI development set).

For GPT3, we expeirment with two baselines, one with 3 demonstration tuples with balanced labels, and the other with four, half of which having the neural label. The latter aims to control for the label distribution shift invited by the data augmentation process. The trend is different than the finetuend T5-large: GPT-3 with data augmentation sees decreased accuracy and increased attentiveness, but still far from perfect attentiveness. These results suggest that this simple augmentation procedure can be used as a method for improving model's robustness, as measured by CAT.

[13]This will increase the data size by the proportion of non-default labels, e.g. the new training set size will increase by 66% in the case of balanced 3-label classification like MNLI.

7 Conclusions

We studied models' reliance on partial inputs and proposed a method to quantify it. We applied our method, Counterfactual Attentiveness Test (CAT) on multiple tasks, datasets, and setups including fine-tuning and in-context learning. We found that the spurious correlations between a part of the inputs and the labels are not always indicative of the full model's reliance on such heuristics in the supervised setup. For instance, we found that while partial inputs in SQuAD2.0 are correlated with the labels, models trained on it are attentive to both parts of the input. This is in contrast to MNLI, whose hypotheses and labels are highly correlated, and models trained on it are also inattentive to the premises. In addition, in ICL, we found that more demonstrations lead to decreased attentiveness in models such as GPT-3.

8 Limitations

While CAT allows automatic evaluation of the attentiveness of a model on the full input, it has some limitations.

Not a Standalone Evaluation CAT is not meant to replace the standard model's performance, but to provide insights into models' behavior. For instance, a model can perform close-to-perfect on CAT if the model relies on lexical overlap heuristics, as the counterfactual pairs are not likely to have high lexical overlap. As such, while the model is likely to perform well on CAT, it will fail on other benchmarks (e.g., HANS, McCoy et al. 2019). In addition, CAT will not produce a meaningful score for models that did not learn the task and always predict the same label. We highlight the importance of reporting both the standard metric used for a task or dataset, as well as ours for a more complete evaluation.

Random Counterfactual Assumption The automatic counterfactual generation process assumes that two random instances from the same dataset will not be related, and as such the label will change. However, this is an assumption that should be manually evaluated. As we discuss in Section 4.2, and in Table 2, we show that for the datasets we considered this assumption is mostly true (with the smallest percentage being 92% correct). However, in preliminary experiments we considered another dataset where this assumption didn't hold in many cases: SNLI (Bowman et al., 2015). Out of 50 counterfactual instances we manually inspected, we found the *neutral* label to be correct only in 70% of the cases. Consider the following counterfactual example we generated from SNLI: "P: A man standing in front of a chalkboard points at a drawing. H: The person is in a hat has a big bag while walking on a tough terrain.". We labeled this example as *contradiction* because both sentences discuss the same entity. This is as opposed to the original *entailment* label this hypothesis was paired with the following premise: "P: A person in a red hat with a huge backpack going hiking.". The neutral label assumption does not hold in the SNLI case is mainly due to the source of data used to collect SNLI, which was based on image captions. In addition, the dataset lexicon is not diverse enough, and we often observe the same entities mentioned in the generated counterfactuals. Inspecting the SNLI gold annotations, it is clear that when an entity is mentioned in both P & H, one must assume that they refer to the same entity if reasonably possible.[14] This result led us to discard SNLI from the analysis, as our assumption was not accurate enough.

SNLI is the only additional dataset we considered for this study and did not include in the analysis due to our assumption not holding in practice. We note though, that when analyzing other datasets, a practitioner should validate the assumption and manually annotate several counterfactual instances before conducting the analysis.

Scope The second limitation is the scope of our method. Since CAT tests the attentiveness of models, and expecting the prediction on the counterfactual to change, we only include instances where the original predictions are non-neutral for computing attentiveness. This leaves cases where a model that achieves a good score on CAT may be inattentive to partial inputs in those cases. Therefore, evidence of attentiveness cannot rule out inattentiveness on neutral instances, and similarly evidence of inattentiveness may be even more serious when originally-neutral instances are included. We leave it to future work to automatically estimate attentiveness for all possible instances.

Acknowledgments

We want to thank Alisa Liu, Sofia Serrano, Marius Mosbach, and Shauli Ravfogel for discussions

[14] Some of the common mentioned entities are: man, woman, boy, people.

and feedback on this project and this draft. This work was supported in part by the National Science Foundation (NSF) grants 2007398, 2217154, and NSF IIS-2046873.

References

Elron Bandel, Yoav Goldberg, and Yanai Elazar. 2022. Lexical generalization improves with larger models and longer training. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4398–4410, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. ArXiv:2210.11416.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, page 177–190, Berlin, Heidelberg. Springer-Verlag.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7659–7666.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Eisenstein. 2022. Informativeness and invariance: Two perspectives on spurious correlations in natural language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4326–4331, Seattle, United States. Association for Computational Linguistics.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Ruth C Fong and Andrea Vedaldi. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pages 3429–3437.

Zee Fryer, Vera Axelrod, Ben Packer, Alex Beutel, Jilin Chen, and Kellie Webster. 2022. Flexible text generation for counterfactual fairness probing. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 209–229.

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.

Matt Gardner, William Merrill, Jesse Dodge, Matthew E Peters, Alexis Ross, Sameer Singh, and Noah A Smith. 2021. Competency problems: On finding and removing artifacts in language data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1801–1813.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. 2022. Inducing causal structure for interpretable neural networks. In *International Conference on Machine Learning*, pages 7324–7338. PMLR.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking nli systems with sentences that require simple lexical inferences. In *Proceedings of the*

56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 650–655.

Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2019. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *Int. J. Comput. Vis.*, 127(4):398–414.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel 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, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Jack Hessel and Lillian Lee. 2020. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 861–877.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. *arXiv preprint arXiv:2305.10160*.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Divyansh Kaushik and Zachary C Lipton. 2018a. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015.

Divyansh Kaushik and Zachary C. Lipton. 2018b. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Gouthaman KV and Anurag Mittal. 2020. Reducing language biases in visual question answering with visually-grounded question encoder. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIII*, volume 12358 of *Lecture Notes in Computer Science*, pages 18–34. Springer.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, 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.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *EMNLP*.

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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al.

2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Judea Pearl. 2009. *Causality*. Cambridge university press.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Anna Rogers. 2023. Closed ai models make bad baselines. Blogpost on Towards Data Science.

Rachel Rudinger, Vered Shwartz, Jena D Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675.

Amrita Saha, Rahul Aralikatte, Mitesh M Khapra, and Karthik Sankaranarayanan. 2018. Duorc: Towards complex language understanding with paraphrased reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1693.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951.

Lakshay Sharma, L. Graesser, Nikita Nangia, and Utku Evci. 2019. Natural language understanding with the quora question pairs dataset. *ArXiv*, abs/1907.01041.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913.

Neha Srikanth and Rachel Rudinger. 2022. Partial-input baselines show that NLI models can ignore context, but they don't. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4753–4763, Seattle, United States. Association for Computational Linguistics.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6418–6428. Association for Computational Linguistics.

Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. 2020. Learning what makes a difference from counterfactual examples and gradient supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 580–599. Springer.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2020. Is multihop qa in dire condition? measuring and reducing disconnected reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8846–8863.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. 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, New Orleans, Louisiana. Association for Computational Linguistics.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving

models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723.

Zhengxuan Wu, Karel D'Oosterlinck, Atticus Geiger, Amir Zur, and Christopher Potts. 2022. Causal proxy models for concept-based model explanations. *arXiv preprint arXiv:2209.14279*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

A Models' Standard Performance

We report the majority-class accuracy and the full results of the supervised models (BERT, RoBERTa, DeBERTa, both versions of T5 and Flan-T5) in Table 8 for NLI and PD. The results of the models used in few-shot in-context learning experiments (both versions of T5, Flan-T5 and GPT-3) are reported in Table 9.

We highlight a few interesting trends. First, DeBERTa-large and Flan-T5-large obtain similar performance across the tasks and datasets. Second, while T5 v1 typically performs better across tasks and models' sizes than v1.1 - which is expected as the training data of many of these dataset were part of training data of the later - it is especially expected noticable on RTE, a very small dataset (2,490 instances for training) - and in the smaller models. For instance, T5-small performs 67.87% on RTE, while the corresponding v1.1 model only achieves 47.29%, which is worse than random.

The ICL results can be found in Appendix A in Tables 8, 10, 9; they illustrate two opposing trends. GPT-3 benefits from more demonstrations, where the biggest benefits happen with three-shots, and additional demonstrations' benefit is more subtle. For instance, GPT-3 achieves 53.7% accuracy on MNLI in zero-shot, then 74.9% with 3-shots and 77.4% with 9-shots. On the other hand, the performance of both the T5 and Flan-T5 models, across all sizes (from small to XL), performs worse the more demonstrations are shown.[15] For instance, on MNLI Flan-T5-XL achieves 89.5% accuracy, slightly decreases to 89.2 with 3-shot, and then completely fails with 6-, and 9-shots (37.8% and 32.3%, respectively). While Min et al. (2022) found that demonstrations contribute by providing examples of the label space, input distribution, and format, and are not as attentive to the content itself, we find that demonstrations harm models in performing the task (except for GPT-3).

In the PD datasets, we notice that the performance of all models are rather high, and the gap between all models is small (5.37 in QQP).

B Partial Input Baseline

We report the full results of the partial input baselines in Tables 12, 13.

In NLI, we train models on the *hypotheses* (or *sentence1* in RTE), *sentence1* in PD, either on the *question* or *paragraph* in RC,[16] and either *image* or the *text* in VLR.

On QQP, T5-base trained on the corresponding dataset achieves 78.81% (compared to the 63.18% majority accuracy).

Similarly, models finetuned on VLR datasets perform much higher than the majority class (24.2%) However, while image-only and text-only models perform similarly when finetuned for NLVR2, we observe that the text-only performance is much higher for VQA v2 dataset (39.9% for BLIP and 42.1% for ViLT) than the image-only ($\sim$ 24%). The image-only model performs close to the majority class baseline. This demonstrates unequal correlations in different modalities.

C Full CAT Results

Here we provide the full results for all datasets we consider. The results are summarized in Table 15.

D Comparable Counterfactuals

In Section 5.1 we report the results of each model based on its own original predictions. As a result, the models evaluated on the same dataset are not entirely comparable since they are evaluated on different subsets. As such, in this section we perform the evaluation on the same subsets across models, to perform a fair comparison. We report the results in Table 16. The average difference between each model between the main results and the comparable ones on the same subset is minimal. This results suggest that even if the models are evaluated on different subsets, it is enough to gain insights about the inspected models.

E Comparison with Srikanth and Rudinger (2022)

We provide additional comparison with Srikanth and Rudinger (2022) in Table 17.

F Responsible Research Checklist

License All datasets we use use the cc-by-3.0 license, besides PAWS, which has a different license: https://github.com/google-research-datasets/paws/blob/master/LICENSE.

Compute To run experiments we used nvidia A-100 gpus, for an estimated time of 100 hours.

[15] Except for Flan-T5 Small that gets worse than random.

[16] Following Kaushik and Lipton (2018b) it is implemented by randomly sampling a different passage from the corpus and adding the answer candidate to this passage in a random location.

	NLI			PD		RC	
Dataset	**RTE**	**MNLI**	**WANLI**	**QQP**	**PAWS**	**SQuAD**	**NewsQA**
BERT-base	65.3	83.5	68.6	91.0	90.7	74.1	59.4
BERT-large	70.0	86.3	69.9	91.5	93.3	80.5	61.8
RoBERTa-base	71.1	86.8	72.7	91.6	94.4	80.3	61.7
RoBERTa-large	79.8	90.1	75.4	92.1	95.7	84.9	65.7
DeBERTa-base	76.5	90.0	75.4	92.4	95.6	87.2	65.1
DeBERTa-large	87.7	91.3	76.7	93.0	95.7	89.6	66.0
T5-base	67.5	85.8	70.2	91.5	93.9	70.8	61.0
T5-large	87.7	89.3	76.2	92.1	94.9	87.5	62.4
T5-1.1-base	45.5	60.9	49.8	87.7	55.8	78.9	59.7
T5-1.1-large	52.7	90.5	71.3	90.3	92.6	88.7	61.7
FLAN-T5-base	79.8	86.1	71.2	91.7	94.0	83.3	60.5
FLAN-T5-large	89.5	90.4	76.4	91.0	95.1	89.4	63.3

Table 8: Standard evaluation performance on respective datasets.

		NLI			PD	
Model	**# Tuples**	**RTE**	**MNLI**	**WANLI**	**QQP**	**PAWS**
Majority	-	52.7	35.4	48.3	63.2	55.8
GPT3	0	74.7	53.7	58.3	50.5	71.7
GPT3	1	86.3	74.9	55.8	80.0	77.6
GPT3	2	86.3	76.3	58.0	81.0	77.3
GPT3	3	88.1	77.4	58.1	81.5	78.1
Flan-T5 Small	0	47.7	38.1	24.3	63.2	55.7
Flan-T5 Small	1	44.4	43.5	35.7	63.3	49.8
Flan-T5 Small	2	44.4	35.4	31.4	63.4	51.4
Flan-T5 Small	3	50.5	34.5	31.5	63.3	54.2
Flan-T5 Base	0	75.5	73.5	48.3	70.8	85.8
Flan-T5 Base	1	74.7	71.4	46.2	68.6	83.4
Flan-T5 Base	2	69.3	36.5	38.1	68.4	79.2
Flan-T5 Base	3	50.5	32.3	34.8	68.2	51.5
Flan-T5 Large	0	88.1	86.8	58.7	84.9	91.3
Flan-T5 Large	1	88.4	86.3	59.1	84.4	90.6
Flan-T5 Large	2	75.8	37.5	40.7	84.5	86.2
Flan-T5 Large	3	50.5	32.1	37.8	84.4	52.5
Flan-T5 XL	0	89.5	89.5	61.7	85.9	94.4
Flan-T5 XL	1	89.2	89.2	62.0	85.6	94.2
Flan-T5 XL	2	76.5	37.8	42.0	85.6	88.8
Flan-T5 XL	3	50.5	32.2	35.1	85.5	53.2
T5 Small	0	50.5	80.4	52.7	88.4	N/A
T5 Small	1	47.3	42.0	35.7	70.3	N/A
T5 Small	2	47.3	38.1	34.2	65.7	N/A
T5 Small	3	47.3	34.5	32.0	63.4	N/A
T5 Base	0	67.5	85.6	59.0	90.6	N/A
T5 Base	1	49.1	40.1	33.2	69.0	N/A
T5 Base	2	51.3	37.9	37.1	66.4	N/A
T5 Base	3	49.5	35.1	36.0	64.4	N/A
T5 Large	0	85.9	89.7	61.7	91.3	N/A
T5 Large	1	82.7	42.7	36.6	74.9	N/A
T5 Large	2	75.8	40.2	33.3	70.9	N/A
T5 Large	3	57.8	35.7	25.7	67.4	N/A
T5 3B	0	87.7	91.1	63.3	90.6	N/A
T5 3B	1	52.0	40.3	31.3	68.6	N/A
T5 3B	2	47.3	36.3	21.2	64.7	N/A
T5 3B	3	43.0	33.7	16.4	63.3	N/A

Table 9: In-context learning dev. set accuracy. T5 models fail to output valid labels on PAWS.

Dataset	VQA v2	NLVR2
BLIP	77.5	81.1
ViLT	70.3	74.6

Table 10: Standard performance on the VLR tasks benchmarked on the finetuned model with full-inputs

Model	#Tuples	NLI			PD	
		RTE	MNLI	WANLI	QQP	PAWS
GPT3	0	$96.3_{\pm 0.9}$	$94.6_{\pm 0.1}$	$94.3_{\pm 0.3}$	$97.2_{\pm 0.1}$	$100.0_{\pm 0.0}$
GPT3	1	$100.0_{\pm 0.0}$	$31.3_{\pm 0.3}$	$36.7_{\pm 0.4}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
GPT3	2	$100.0_{\pm 0.0}$	$26.0_{\pm 0.7}$	$41.0_{\pm 0.8}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
GPT3	3	$100.0_{\pm 0.0}$	$25.6_{\pm 0.5}$	$41.8_{\pm 0.9}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
Flan-T5-base	0	$100.0_{\pm 0.0}$	$55.9_{\pm 0.2}$	$57.9_{\pm 0.3}$	$99.7_{\pm 0.0}$	$99.9_{\pm 0.0}$
Flan-T5-base	1	$98.6_{\pm 1.1}$	$23.5_{\pm 0.7}$	$23.9_{\pm 0.6}$	$99.7_{\pm 0.0}$	$99.9_{\pm 0.0}$
Flan-T5-large	0	$100.0_{\pm 0.0}$	$94.1_{\pm 0.1}$	$94.3_{\pm 0.3}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
Flan-T5-large	1	$99.9_{\pm 0.2}$	$95.8_{\pm 0.6}$	$96.5_{\pm 0.2}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
Flan-T5-XL	0	$100.0_{\pm 0.0}$	$99.0_{\pm 0.1}$	$99.2_{\pm 0.1}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
Flan-T5-XL	1	$100.0_{\pm 0.0}$	$98.5_{\pm 0.5}$	$99.3_{\pm 0.1}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
T5-base	0	$100.0_{\pm 0.0}$	$72.6_{\pm 0.2}$	$87.6_{\pm 0.3}$	$99.9_{\pm 0.0}$	-
T5-base	1	$99.8_{\pm 0.3}$	$53.7_{\pm 2.6}$	$57.6_{\pm 7.8}$	$97.9_{\pm 0.4}$	-
T5-large	0	$100.0_{\pm 0.0}$	$65.9_{\pm 0.2}$	$77.6_{\pm 0.3}$	$100.0_{\pm 0.0}$	-
T5-large	1	$99.0_{\pm 0.8}$	$43.0_{\pm 4.0}$	$37.2_{\pm 8.9}$	$98.8_{\pm 0.3}$	-
T5-3B	0	$100.0_{\pm 0.0}$	$61.7_{\pm 0.2}$	$72.7_{\pm 0.4}$	$100.0_{\pm 0.0}$	-
T5-3B	1	$100.0_{\pm 0.0}$	$27.3_{\pm 3.3}$	$28.6_{\pm 1.9}$	$98.8_{\pm 0.4}$	-

Table 11: Accuracy of in-context learning models on the standard dev. split without perturbation, averaged across five different random seeds. T5 models fail to output valid labels on PAWS. The small-sized T5 model, in the three-shot setting, always makes neutral predictions on the original RTE dev. set, making its counterfactual accuracy unavailable.

Setup	Model/Data	VQA v2	NLVR2
Image-only	BLIP	24.2	50.9
	ViLT	25.4	49.3
Text-only	BLIP	40.0	42.2
	ViLT	50.7	49.3

Table 12: Standard performance on the Visual Reasoning tasks benchmarked on partial inputs.

	NLI			PD		RC - Question Only			RC - Paragraph Only		
Model/Data	**RTE**	**MNLI**	**WANLI**	**QQP**	**PAWS**	**SQuAD**	**DuoRC**	**NewsQA**	**SQuAD**	**DuoRC**	**NewsQA**
Majority	52.7	35.4	48.3	63.2	55.8	33.4		28.8	33.4		28.8
BERT-base	56.0	56.6	50.8	77.9	56.6	95.6	67.2	90.7	50.1	6.1	27.5
BERT-large	57.4	57.1	50.7	77.5	56.2	95.5	69.7	91.8	50.1	6.3	27.6
RoBERTa-base	52.0	59.3	53.3	77.9	58.1	95.1	70.2	96.3	50.1	6.6	27.5
RoBERTa-large	56.3	59.7	51.7	77.7	55.8	96.5	72.2	96.7	50.1	6.8	27.5
DeBERTa-base	61.4	60.0	52.7	76.5	58.0	98.0	70.9	96.6	50.1	7.0	27.5
DeBERTa-large	61.4	61.1	52.4	78.3	58.4	98.5	73.2	96.8	50.1	7.0	27.6
T5-base	56.0	57.9	50.1	78.8	58.6	93.5	54.0	61.2	50.2	5.9	27.5
T5-large	63.2	61.2	54.7	77.8	56.9	97.1	58.4	64.3	50.1	6.9	27.5
T5-1.1-base	56.3	56.4	49.0	77.9	56.1	95.1	54.8	62.3	50.1	5.9	27.5
T5-1.1-large	53.1	61.7	49.3	74.3	53.9	97.7	58.5	64.1	50.1	6.8	27.5
FLAN-T5-base	53.8	57.6	50.9	78.6	58.0	95.2	56.4	62.0	50.1	6.5	27.5
FLAN-T5-large	59.2	61.5	55.7	75.5	56.3	97.5	58.5	64.4	50.1	7.0	27.5

Table 13: Partial-input baselines on NLI, PD, and RC.

Task	**Instruction**
RTE	You're given a pair of sentences: a Text and a Hypothesis. Your job is to determine the relation between them based on your inference from the statement and your commonsense knowledge. Answer 'Entailment' if the Hypothesis can be inferred from the Text; Answer 'Not entailment' if the Hypothesis disagrees with the Text.
MNLI, WANLI	You're given a pair of sentences: a Premise and a Hypothesis. Your job is to determine the relation between them based on your inference from the statement and your commonsense knowledge. Answer 'Entailment' if the Hypothesis can be inferred from the Premise; Answer 'Contradiction' if the Hypothesis disagrees with the Premise Answer 'Neutral' if the Hypothesis can neither be inferred from the Premise nor disagrees with the Premise.
QQP	You're given a pair of questions. Your job is to determine whether they are semantically equivalent. Answer 'Paraphrase if they bear the same meaning; Answer 'Not paraphrase' if they have different meanings.
PAWS	You're given a pair of sentences. Your job is to determine whether they are semantically equivalent. Answer 'Paraphrase' if they bear the same meaning; Answer 'Not paraphrase' if they have different meanings.

Table 14: Textual instructions for the instruction-fintuned models in the in-context learning experiments.

	NLI			PD		RC		
Dataset	**RTE**	**MNLI**	**WANLI**	**QQP**	**PAWS**	**SQuAD**	**DuoRC**	**NewsQA**
BERT-base	$99.8_{\pm 0.3}$	$74.6_{\pm 0.3}$	$97.8_{\pm 0.2}$	$99.9_{\pm 0.0}$	$99.3_{\pm 0.2}$	$98.0_{\pm 0.0}$	$97.4_{\pm 0.1}$	$99.0_{\pm 0.1}$
BERT-large	$100.0_{\pm 0.0}$	$64.3_{\pm 0.1}$	$99.8_{\pm 0.1}$	$99.9_{\pm 0.0}$	$85.9_{\pm 0.3}$	$98.0_{\pm 0.2}$	$97.6_{\pm 0.1}$	$99.1_{\pm 0.1}$
RoBERTa-base	$99.8_{\pm 0.3}$	$73.4_{\pm 0.5}$	$99.5_{\pm 0.1}$	$99.9_{\pm 0.0}$	$80.5_{\pm 0.6}$	$98.3_{\pm 0.1}$	$97.7_{\pm 0.1}$	$99.1_{\pm 0.2}$
RoBERTa-large	$100.0_{\pm 0.0}$	$69.8_{\pm 0.2}$	$99.6_{\pm 0.1}$	$99.9_{\pm 0.0}$	$78.7_{\pm 0.3}$	$98.2_{\pm 0.2}$	$98.1_{\pm 0.1}$	$99.1_{\pm 0.1}$
DeBERTa-base	$97.0_{\pm 1.6}$	$78.8_{\pm 0.4}$	$99.5_{\pm 0.1}$	$99.9_{\pm 0.0}$	$99.9_{\pm 0.1}$	$97.6_{\pm 0.1}$	$97.6_{\pm 0.3}$	$99.0_{\pm 0.1}$
DeBERTa-large	$100.0_{\pm 0.0}$	$68.6_{\pm 0.3}$	$99.4_{\pm 0.1}$	$99.9_{\pm 0.0}$	$99.1_{\pm 0.1}$	$98.0_{\pm 0.2}$	$97.9_{\pm 0.1}$	$98.9_{\pm 0.1}$
T5-base	$99.4_{\pm 0.7}$	$81.3_{\pm 0.1}$	$99.0_{\pm 0.2}$	$99.9_{\pm 0.0}$	$99.0_{\pm 0.1}$	$99.1_{\pm 0.1}$	$96.5_{\pm 0.1}$	$99.5_{\pm 0.0}$
T5-large	$100.0_{\pm 0.0}$	$71.5_{\pm 0.4}$	$99.8_{\pm 0.1}$	$99.9_{\pm 0.0}$	$100.0_{\pm 0.0}$	$99.5_{\pm 0.1}$	$97.8_{\pm 0.2}$	$99.7_{\pm 0.0}$
T5-1.1-base	-	$46.9_{\pm 0.6}$	-	$97.1_{\pm 0.1}$	-	$99.4_{\pm 0.0}$	$96.9_{\pm 0.1}$	$99.0_{\pm 0.1}$
T5-1.1-large	-	$79.4_{\pm 0.3}$	$99.8_{\pm 0.1}$	$99.9_{\pm 0.0}$	-	$99.4_{\pm 0.1}$	$97.6_{\pm 0.0}$	$99.6_{\pm 0.1}$
FLAN-T5-base	$99.9_{\pm 0.3}$	$81.8_{\pm 0.3}$	$99.5_{\pm 0.1}$	$99.9_{\pm 0.0}$	$98.3_{\pm 0.2}$	$99.4_{\pm 0.1}$	$97.4_{\pm 0.0}$	$99.4_{\pm 0.2}$
FLAN-T5-large	$100.0_{\pm 0.0}$	$77.1_{\pm 0.2}$	$99.8_{\pm 0.1}$	$99.9_{\pm 0.1}$	$100.0_{\pm 0.1}$	$99.5_{\pm 0.1}$	$98.2_{\pm 0.1}$	$99.7_{\pm 0.0}$

Table 15: CAT results on natural language inference (NLI), paraphrase detection (PD) and reading comprehension (RC). Results are calculated by computing the mean attentiveness of five counterfactuals per instance, and their standard deviation. Higher numbers indicate better model attentiveness. We mark results for models whose standard performance is not significantly better than random with '-'.

Dataset	RTE	MNLI	WANLI
BERT-base	93.3	70.8	98.4
BERT-large	100.0	60.2	100.0
RoBERTa-base	100.0	68.8	99.9
RoBERTa-large	100.0	64.7	99.4
DEBERTA-base	100.0	74.7	99.4
DeBERTa-large	100.0	64.5	99.3
T5-base	93.3	77.5	98.9
T5-large	100.0	67.2	100.0
T5-1.1-base	-	43.3	-
T5-1.1-large	-	74.6	100.0
FLAN-T5-base	100.0	78.3	99.4
FLAN-T5-large	100.0	72.8	100.0

Table 16: Comparable partial-inputs counterfactuals on NLI datasets.

Property	Srikanth and Rudinger (2022)	Our Work
Tasks	NLI	NLI, PD, RC, V&L-R
NLI-Datasets	SNLI, δ-NLI	RTE, MNLI, WANLI
Models	RoBERTa-base	BERT*, RoBERTa*, DeBERTa*, T5* T5-v1.1*, Flan-T5*, GPT-3
Counterfactuals	Manual	Automatic
Examples subset	Heuristic prone	Non-neutral prediction
Epochs	2	10
New labels	All	Neutral
Conclusions	No reliance	Data & Model dependent

Table 17: Summarizing our setup compared to Srikanth and Rudinger (2022). * signifies we test several model sizes from the same family.