How Often Are Errors in Natural Language Reasoning Due to Paraphrastic Variability?

Neha Srikanth
Computer Science
University of Maryland, USA
nehasrik@umd.edu

Marine Carpuat
Computer Science
University of Maryland, USA
marine@cs.umd.edu

Rachel Rudinger
Computer Science
University of Maryland, USA
rudinger@umd.edu

Abstract

Large language models have been shown to behave inconsistently in response to meaning-preserving paraphrastic inputs. At the same time, researchers evaluate the knowledge and reasoning abilities of these models with test evaluations that do not disaggregate the effect of paraphrastic variability on performance. We propose a metric, P_C, for evaluating the *paraphrastic consistency* of natural language reasoning models based on the probability of a model achieving the same correctness on two paraphrases of the same problem. We mathematically connect this metric to the proportion of a model's variance in correctness attributable to paraphrasing. To estimate P_C, we collect ParaNlu, a dataset of 7,782 human-written and validated paraphrased reasoning problems constructed on top of existing benchmark datasets for defeasible and abductive natural language inference.[1] Using ParaNlu, we measure the paraphrastic consistency of several model classes and show that consistency dramatically increases with pretraining but not fine-tuning. All models tested exhibited room for improvement in paraphrastic consistency.

1 Introduction

The NLP community has transitioned away from ''deeper'' abstract semantic representations (e.g., FrameNet (Baker et al., 1998)) towards ''shallower'' representations (e.g., Universal Dependencies (Nivre et al., 2016)) which retain attributes of their original surface form. The culmination of this trend is to use natural language as a semantic representation itself for evaluating a model's reasoning ability. This has enabled rapid advancement across a host of tasks including NLI (Bowman et al., 2015) and QA (Rajpurkar et al., 2016), with the latest generation of large language models saturating many benchmark natural language understanding datasets. However, natural language as a meaning representation is highly ambiguous (Schubert, 2015). While versatile and compact, it leaves open the possibility that systems are not robust to *different ways* of expressing the same meaning in natural language.

Benchmark evaluation datasets such as SNLI (Bowman et al., 2015) consist of a collection of reasoning problems designed to probe particular aspects of commonsense knowledge, with each example represented by a *singular* linguistic expression. When a system gets a particular example correct, it is only evidence that it was able to correctly reason *for the particular phrasing used in the example*, allowing for the possibility of systems that can correctly solve one form of a reasoning problem, but not others. Conversely, if a model gets a question wrong, how can we tell if the error was due to a failure in language understanding or a failure in reasoning?

Consider the defeasible reasoning example in Figure 1. A language model finetuned on the δ-NLI dataset (Rudinger et al., 2020) may correctly predict that the original update sentence *strengthens* a human's belief in the hypothesis sentence. However, different linguistic expressions of that same update sentence may yield high variance in a model's predictions. If models stay *consistent* in the face of paraphrastic variability, we may conclude that correctly reasoning about one expression is indicative of an understanding of that *reasoning problem*, a desirable property of teaching machines to reason entirely in natural language.

We explore the sensitivity of natural language reasoning models to paraphrasing as a way to better characterize their knowledge and reasoning ability, contextualize their performance on evaluation sets, and evaluate room for improvement on the basis of consistency. Under the assumption

[1] We publicly release all data and code at https://github.com/nehasrikn/paraphrase-nlu.

Transactions of the Association for Computational Linguistics, vol. 12, pp. 1143–1162, 2024. https://doi.org/10.1162/tacl_a_00692
Action Editor: Jacob Eisenstein. Submission batch: 12/2023; Revision batch: 5/2024; Published 9/2024.

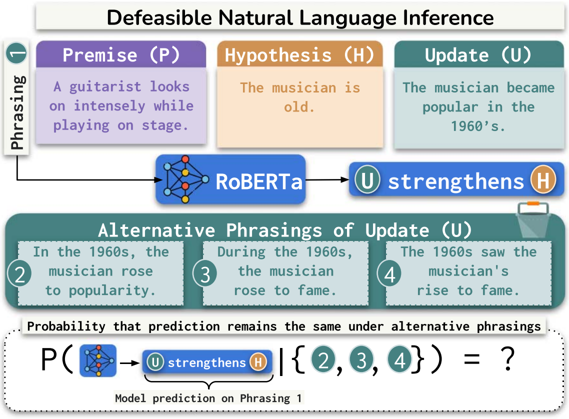

Defeasible Natural Language Inference
Phrasing
Premise (P)
A guitarist looks on intensely while playing on stage.
Hypothesis (H)
The musician is old.
Update (U)
The musician became popular in the 1960's.
RoBERTa
U strengthens H
Alternative Phrasings of Update (U)
2 In the 1960s, the musician rose to popularity.
3 During the 1960s, the musician rose to fame.
4 The 1960s saw the musician's rise to fame.
Probability that prediction remains the same under alternative phrasings
P(→ U strengthens H |{2, 3, 4}) = ?
Model prediction on Phrasing 1

Figure 1: δ-NLI instance with a set of paraphrased update sentences. We study **paraphrastic consistency**, or the probability that a model's prediction for two phrasings of the same problem match.

that world and linguistic knowledge are separable, our study attempts to disentangle the two by generating examples that hold the required world knowledge of a reasoning problem constant while modifying its surface form, a problem formulation with connections to causality (Stolfo et al., 2023) and counterfactual invariance (Veitch et al., 2021; Kaushik et al., 2020).

To study this, we build on top of two NLU datasets—Abductive NLI (Bhagavatula et al., 2019) and Defeasible NLI (Rudinger et al., 2020)—by collecting paraphrases of reasoning problems using *label-preserving* paraphrasing, a functional change of traditional paraphrasing that preserves the semantics of the *core reasoning problem*. Our dataset, ParaNlu, contains 7,782 human-elicited and 7,295 model-elicited paraphrases across 1,000 reasoning problems spanning both datasets. We select diverse examples to paraphrase ranging in difficulty (Sakaguchi et al., 2021) and model confidence. Our dataset is *entirely manually validated*, ensuring semantic equivalence while maximizing paraphrase diversity.

We measure **paraphrastic consistency** (P_C), or the likelihood of model's prediction remaining consistent under different phrasings, in order to understand the types of surface-form changes that models are sensitive to. We study the relationship between consistency and various data conditions and modeling paradigms, exploring factors such as data source, example difficulty, model complexity, and training dynamics. *For their given accuracy level*, we find that models still have room to improve on paraphrastic consistency. Since no model demonstrates high accuracy *and* high paraphrastic consistency, we conclude that attempts to measure their *reasoning* abilities will be confounded by inconsistencies in their *linguistic* abilities.

2 Paraphrastic Consistency

In principle, natural language reasoning tasks like abductive NLI and defeasible NLI require the ability to linguistically *decode* the meaning of the text that expresses an underlying problem, as well as the knowledge and reasoning capabilities to *solve* the underlying problem.

By analogy, consider evaluating a child's understanding of the concept of addition. Instead of simply presenting them with a mathematical expression (say, $7+7$), we write a *word problem* that can be answered by (1) understanding the situation in natural language, (2) recognizing that the answer corresponds to the mathematical reasoning problem $7+7$, and finally, (3) solving $7+7$. If the child answers incorrectly, we must figure out whether they did not understand the goal of the word problem or were not able to perform the arithmetic in order to evaluate their mathematical reasoning ability.

For models tasked with natural language reasoning problems, teasing apart these two failure modes (namely, deficiencies in language understanding versus deficiencies in knowledge or reasoning) requires more than reporting test set *accuracy*. The design of natural language reasoning test sets does not facilitate this type of analysis: If a test set contains 100 different natural language reasoning problems, and a model correctly answers 80% of them, which failure mode should we attribute the 20% of errors to?

For practitioners, it is useful to characterize performance by measuring *paraphrastic consistency* alongside accuracy: How likely is it that a model's prediction for a natural language reasoning problem will remain the same given a *different phrasing* of the problem? We collect a dataset that changes the *language* of NLI examples while maintaining the *underlying logic* of the problem to tease the two apart. For a test example, x, we collect a set of paraphrases $\{x'_1, x'_2, ...\}$, which we call a *bucket* (🪣). After collecting many such buckets, we can directly estimate the probability

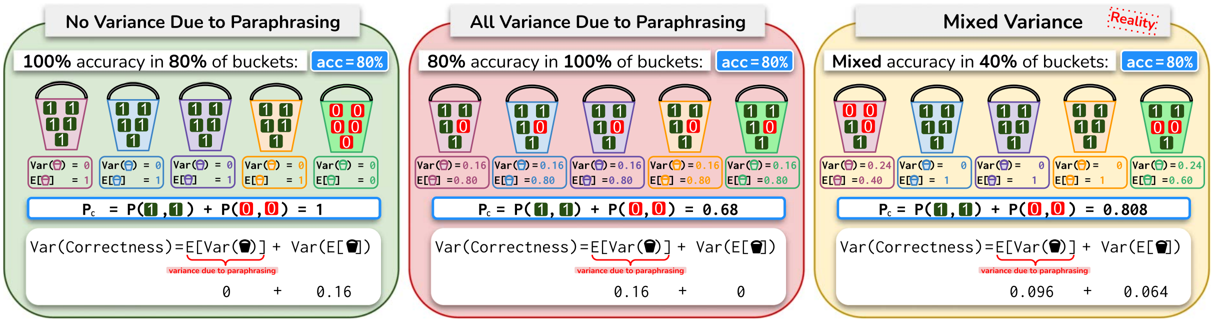

No Variance Due to Paraphrasing
100% accuracy in 80% of buckets:
acc=80%
Pc = P(1,1) + P(0,0) = 1
Var(Correctness)=E[Var(🪣)] + Var(E[🪣])
variance due to paraphrasing
0 + 0.16
All Variance Due to Paraphrasing
80% accuracy in 100% of buckets:
acc=80%
Pc = P(1,1) + P(0,0) = 0.68
Var(Correctness)=E[Var(🪣)] + Var(E[🪣])
variance due to paraphrasing
0.16 + 0
Mixed Variance
Reality
Mixed accuracy in 40% of buckets:
acc=80%
Pc = P(1,1) + P(0,0) = 0.808
Var(Correctness)=E[Var(🪣)] + Var(E[🪣])
variance due to paraphrasing
0.096 + 0.064

Figure 2: Three scenarios, all with equivalent overall accuracy of 80%, illustrating different distributions of variance in model predictions leading to different P_C values. Buckets represent underlying commonsense reasoning problems. Numbers within buckets represent model correctness on 5 paraphrases. Most models achieve a mix of accuracy within and across buckets of paraphrased examples.

that a model's prediction for any two paraphrases belonging to the same bucket will be the same.

2.1 Measuring Paraphrastic Consistency

When authoring a test example for a natural language reasoning task, a crowdworker has many linguistic choices for expressing the underlying reasoning problem. If the purpose of the resulting test set is to evaluate a model's ability to perform the underlying reasoning task, then ideally the crowdworker's choice of phrasing would have no effect on the model's performance. In practice, however, it is known that language models exhibit some degree of sensitivity to paraphrastic variation (Verma et al., 2023; Jiang et al., 2021). To quantify this effect, we pose the following counterfactual question: **If a given test question had been written differently, what is the probability that a model would still receive the same credit for its prediction on the paraphrased question?**

Quantitatively, we introduce a metric of *paraphrastic consistency*, P_C, defined as the probability that a model's predictions for two paraphrases of the same problem, x'_i and x'_j, are either both correct or both incorrect (provided the ground truth labels of x'_i and x'_j match). To formalize P_C, we define the following terms:

- $R_{🪣}$: a discrete random binary variable indicating whether a model's prediction of a paraphrased reasoning problem, x', is correctly predicted (1) or incorrectly predicted (0).
- $\theta_{🪣}$: $\mathbb{E}[R_{🪣}]$, or the average correctness (i.e., accuracy) of the paraphrased problems $(x'_1, x'_2, ...)$ in a particular bucket.
- A: Overall accuracy of model M over a set of natural language reasoning problems. This is equivalent to $\mathbb{E}[\theta_{🪣}]$ across all buckets.

For a binary classification task, where $y \in 0, 1$ denotes the gold label, we then define P_C as:

$$P_C = \underbrace{P(M(x_i) = y, M(x_j) = y)}_{\text{prob. of both predictions correct}} + \underbrace{P(M(x_i) \neq y, M(x_j) \neq y)}_{\text{prob. of both predictions incorrect}} \quad (1)$$

P_C can be estimated from the accuracies of paraphrase buckets, $\theta_{🪣}$, as follows:

$$P_C = \mathbb{E}[\theta_{🪣}^2] + \mathbb{E}[(1 - \theta_{🪣})^2] \quad (2)$$

Informally, the *paraphrastic consistency* of a natural language reasoning model is its ability (or lack thereof) to make the same predictions on paraphrased inputs that, in principle, should yield the same answer. We note that the metric P_C *cannot be computed from a standard test set containing only one phrasing per test example*; to estimate P_C we collect paraphrases of test examples, as described in §4.

Figure 2 illustrates the computation of P_C for different patterns of model behavior. If $P_C = 1$, the model is either entirely correct or incorrect on all paraphrases of the same problem, as in the left-most panel of Figure 2 where each bucket contains only 1's or only 0's. In this case, no errors can be attributed to paraphrastic variance. The minimum P_C occurs when every paraphrase bucket has the same accuracy, as shown in the middle panel; in this case *all* errors are likely due to paraphrastic variability. In practice, P_C lies between these two extremes and *some*, but not all, errors are due to paraphrasing, as in the right-most panel. Modern NLP evaluation sets usually consist of a collection of independent reasoning problems each represented by singular natural language expression, and practitioners often make claims about the reasoning capabilities or world knowledge given a model's accuracy on such evaluation sets. However, as depicted in Figure 2, accuracy presents an incomplete picture of performance: In all three scenarios, the overall accuracy remain 80%, but only the first scenario, in which the model makes equivalent predictions given many alternate phrasings of a reasoning problem, results in a high P_C.

P_C in Practice. P_C can be interpreted as the probability that given two phrasings of the same reasoning problem, the model's two predictions are either both correct or both incorrect. This summary metric allows us to capture the *reliability* of a model's prediction: How confident can we be that a model's prediction would remain correct (or incorrect) if it had been phrased differently? While P_C is lower-bounded by a function of accuracy, we design it as a metric *complementary* to accuracy in order to better characterize model performance, diagnose modeling errors, and benchmark the linguistic reasoning capabilities. For example, when two models achieve similar accuracies, computing their respective paraphrastic consistencies may help as an additional point of comparison. Just as desired model accuracy may be dependent on the application in which it is deployed, different settings may mandate varying appropriate P_C values and practitioners must decide on tolerable thresholds based on their use case.

2.2 Proportion of Variance Attributable to Paraphrasing

Decomposing the *total variance in correctness* across all paraphrased examples in all buckets gives us a clearer picture of the two failure modes described in the opening of §2. Using the law of total variance (Weiss et al., 2006), we decompose the total variance of the correctness across all *paraphrased* examples (the variance of R for all paraphrased problems in all buckets) into two terms: the average variance of correctness within a bucket, $\mathbb{E}[\text{Var}(R_{🪣})]$, and the variance in mean correctness (accuracy) across buckets, $\text{Var}(\mathbb{E}[(R_{🪣})])$, or simply, $\text{Var}(\theta_{🪣})$.

$$\text{Var}(R) = \underbrace{\mathbb{E}[\text{Var}(R_{🪣})]}_{\text{variance from paraphrasing}} + \text{Var}(\theta_{🪣}) \quad (3)$$

This breakdown allows us to better identify the *source* of the variance in correctness. The first term, commonly known as *unexplained variance*, measures <u>variance</u> <u>attributable</u> to <u>paraphrasing</u> (henceforth denoted as VAP) within a bucket. The second, commonly known as *explained variance*, represents variance *across buckets* due to inherent differences in latent characteristics (e.g., difficulty) of different problems. If paraphrasing has no effect, the variance of correctness in each bucket ($\text{Var}(R_{🪣})$) is zero and consequently, the VAP should also be zero. Most evaluation paradigms cannot directly measure VAP, since during data collection, multiple surface forms of the same reasoning problem are not collected. We replicate the conditions under which original examples were produced to simulate different linguistic expressions annotators *could* have chosen.

Mathematical manipulation yields:

$$\begin{aligned} P_C &= 1 - \underbrace{P(M(x_i) = y, M(x_j) \neq y)}_{\text{prediction flips to incorrect}} \\ &\quad - \underbrace{P(M(x_i) \neq y, M(x_j) = y)}_{\text{prediction flips to correct}} \quad (4) \\ P_C &= 1 - \underbrace{\mathbb{E}[\theta_{🪣} \cdot (1 - \theta_{🪣})]}_{\text{prediction flips to incorrect}} \\ &\quad - \underbrace{\mathbb{E}[(1 - \theta_{🪣}) \cdot \theta_{🪣}]}_{\text{prediction flips to correct}} \quad (5) \\ P_C &= 1 - 2 \cdot \mathbb{E}[\theta_{🪣} \cdot (1 - \theta_{🪣})] \quad (6) \\ &= 1 - 2 \cdot \underbrace{\mathbb{E}[\text{Var}(R_{🪣})]}_{\text{VAP}} \quad (7) \end{aligned}$$

Although the primary metric we report throughout the paper is P_C, Equation 7 highlights the direct negative relationship between P_C and variance attributable to paraphrasing (VAP).

Context (C)	Original Target (T)	Admissible Paraphrase	Inadmissible Paraphrase
O1: My cell phone lanyard broke on Wednesday **O2:** My wife was able to repair the lanyard so I did not have to buy one.	h^+: It looked like it was ripped or torn. h^-: It didn't look like it was ripped or torn.	h^+: It seemed to be split and frayed. h^-: It did not appear to be damaged.	h^+: The screen had wear and tear. h^-: It looked like it was in good shape, but apparently it wasn't.
O1: Mike was graduating high school later in the day. **O2:** When he got back home, Mike regretted skipping graduation.	h^+: Mike's parents said they couldn't go, but they did. h^-: Mikes parents didn't show up.	h^+: Despite Mike's parents' claims that they couldn't, they went. h^-: Mike's parents were absent from the graduation ceremony.	h^+: His parents mentioned to him they couldn't go. h^-: Nobody showed up to see Mike's parents.
P: PersonX stops eating fast food. **H:** PersonX is seen as better.	U : PersonX's high school friends think this is strange.	U : This was considered weird by PersonX's friends from high school.	U : PersonX's alumni from high school find this to be peculiar.
P: Two men walking down the sidewalk carrying skateboards. **H:** Two people are walking to the skate park.	U : Both men are adjusting their wheels as they walk.	U : As they stride along, both men are changing the settings on their wheels.	U : While walking the men re-wax their boards.
H: It's okay to cancel an job interview.	U : You are too lazy to get out of bed.	U : You lack the motivation to leave the comfort of your bed.	U : You're feeling sick and can't make yourself rise from bed.

Table 1: Examples of paraphrased abductive NLI (rows 1 and 2) and defeasible NLI (rows 3 – 5) problems in ParaNlu. We show paraphrases written by crowdworkers that are admissible under our definition of label-preserving paraphrasing, and rejected paraphrases that did not meet the label-preserving criteria.

Related to vap is the *proportion* of total variance attributable to paraphrasing (pvap), which is simply vap divided by the total variance:

$$\text{PVAP} = \text{VAP}/\text{Var}(R) \tag{8}$$

Model Confidence and P_C. We choose to characterize paraphrastic consistency via variance in *correctness* instead of variance in *model confidence*. Confidence represents a model's *overall* estimate in the correct answer, and thus conflates confidence in linguistic understanding and problem solving. A low confidence in the *correct label* may indicate that the model understood what the problem was asking, but was unable to reach the answer (or vice versa). Models trained to optimize for *accuracy* are not calibrated to explicitly encode confidence in linguistic decoding or problem-solving ability.

3 Reasoning Tasks

We study paraphrastic consistency across two commonsense reasoning tasks: defeasible reasoning (§3.1) and abductive reasoning (§3.2).

3.1 Defeasible Reasoning

Defeasible reasoning is a mode of reasoning in which inferences or conclusions may be altered or withdrawn in light of new evidence (Reiter, 1980). For example, given the context *''A group of people perform on a stage''*, the natural conclusion *''A group performs for an audience''* may be weakened upon learning that the group is at a rehearsal.

To study defeasible reasoning in language models, Rudinger et al. (2020) introduce the task of defeasible natural language inference (NLI). Traditionally, NLI involves determining whether a **premise** P entails, contradicts, or is neutral in relation to a **hypothesis** H (Giampiccolo et al., 2007). When the premise P and hypothesis H are neutral in relation to one another, defeasible NLI studies whether a third **update** (U) sentence strengthens or weakens H. Namely, a human may determine H more likely to be true when U is a strengthener, and less likely to be true when U is a weakener.

Premise: A woman in shorts throwing a bowling ball down a bowling alley.

Hypothesis: A woman is getting a strike!

Update: The bowling ball falls in the gutter first.

Update Type: Weakener

Rudinger et al. (2020) also introduce δ-NLI, a dataset that extends three existing natural language datasets: SNLI (Bowman et al., 2015), SOCIAL-CHEM-101 (Forbes et al., 2020), and ATOMIC (Sap et al., 2019). For each premise-hypothesis pair (or just hypothesis, in the case of δ-SOCIAL), crowdworkers write multiple strengthening and weakening updates, ensuring a balance. Defeasible NLI is a binary classification task that involves predicting whether the update sentence is a strengthener or a weakener (e.g., the original update in Row 5 of Table 1 is a weakener).

Adopting the terminology from Srikanth and Rudinger (2022), we distinguish between *context* parts of examples, and *target* parts of examples, with our study of consistency concerning *target* portions of examples. For δ-NLI, we consider the premise P and hypothesis H as *context* sentences, and the update U sentence as the target (Table 1).

3.2 Abductive Reasoning

Abduction is inference to the most likely explanation (Peirce, 1974). Such inferences are hypotheses that can best fit one or more incomplete observations.

Observation 1: George decided to buy a TV.
Observation 2: It turned out just as he'd hoped.
Hypothesis $-$: The TV had a cracked screen.
Hypothesis $+$: Upon taking it home and unpacking it, he placed it where he wanted it.

Given the example above, any human would most likely infer the second hypothesis over the first to explain the two observations. Bhagavatula et al. (2019) introduce abductive NLI, an abductive reasoning task formulated binary classification. α-NLI examples consist of two observations O_1 and O_2 (where O_2 occurs some point in time after O_1), h^+, a plausible hypothesis, and h^-, an implausible hypothesis. Given both observations, the task is to determine which hypothesis is more plausible. We treat O_1 and O_2 as context (C) and h^+ and h^- as the target (T) portion of the example.

4 Constructing ParaNlu

We study paraphrastic consistency by paraphrasing the target portions (T) of examples from α-NLI and all three data sources in δ-NLI (δ-SNLI, δ-ATOMIC, and δ-SOCIAL). For an original example x, we collect a *bucket* of paraphrased examples x_i such that the context portions C and gold label l remain identical, while the target portions T are rewritten as *label-preserving paraphrases*. This allows us to modify the surface form of the example while retaining the underlying commonsense reasoning problem.

Label-preserving Paraphrases. We construct quasi-paraphrases (Bhagat and Hovy, 2013) of *target* sentences in reasoning problems by loosening the requirement of semantic equivalence. Given a natural language reasoning task L, a textual instance x of task L with label $\ell_L(x)$, we say that x' is a *label-preserving* paraphrase of x if:

1. $\ell_L(x) = \ell_L(x')$.
2. x' does not contradict any context (C) in x.
3. x' remains consistent with the situation evoked by x.

Label-preserving paraphrases are *functionally* equivalent: Target sentences (hypotheses in α-NLI and updates in δ-NLI) may introduce small bits of information as long as the same scenario is plausibly described. Consider the following δ-NLI example: **P:** *A man stands in front of a cashier and kiosk at a grocery store* **H:** *He is smiling* **U:** *"The man got a discount."* While the sentence *"The man saved 10% with a coupon"* is not semantically equivalent to the strengthening update sentence, it is a valid, label-preserving paraphrase since it retains the logic of the problem and the label. Table 1 shows examples of admissible and inadmissible paraphrases under our definition. Label-preserving paraphrases represent alternative, but equivalent expressions annotators *could* have chosen when writing the original problem that employ similar world knowledge. This is different from *label-altering* edits proposed by Gardner et al. (2020), where minimal human edits that shift the target label were used to create examples for measuring linguistic robustness.

We describe our example selection process for annotation (§4.1) and crowdsourcing protocol for collecting paraphrases and summary statistics of our dataset (§4.2).

4.1 Original Example Selection

We adopt a stratified sampling strategy to obtain diverse examples for annotation that vary in difficulty. To obtain such examples, we leverage AFLite (Sakaguchi et al., 2021), an adversarial filtering (Zellers et al., 2018) algorithm designed to partition datasets based on difficulty using pre-computed dense embeddings of examples fed into an ensemble of n logistic regression classifiers. At each iteration of AFLite, members of the ensemble are trained on random partitions of m examples and evaluated on the remaining validation examples. Each validation example is assigned a score, computed by the proportion of correct predictions. The top-k examples with scores above a threshold τ are subsequently added to the easy partition of the dataset and filtered out, and the process repeats until less than k instances are removed in a particular iteration. Including both types of examples, easy and hard, ensures that PARANLU can support analysis that investigates whether models are inconsistent only on certain *classes* of examples (e.g., those filtered out due to lexical artifacts).

Defeasible NLI. We use the train splits of δ-SNLI, δ-ATOMIC, and δ-SOCIAL to source examples for annotation, since they are large enough to meaningfully partition using AFLITE. We partition each train set into 3 sections: (1) examples to finetune the ROBERTA-base models used to embed examples (ROBERTA$_{\text{embed}}$), (2) training examples for models used for consistency analysis (ROBERTA$_{\text{analysis}}$), and (3) a pool from which examples are sampled for annotation. We partition at the premise-hypothesis (P-H) level to avoid leakage, since multiple examples may contain the same P-H pair, but different updates (U).

For each dataset, we pre-compute example embeddings with the ROBERTA$_{\text{embed}}$ model, and run AFLITE with the $n = 64$ linear classifiers, $\tau = 0.75$, $k = 500$, and $m = 5000$.[2] Examples in the annotation pool with $\tau \leq 0.75$ are added to the *easy* subset, and the those with $\tau > 0.75$ are labeled as *difficult*. We then finetune a separate ROBERTA-large model on ROBERTA$_{\text{analysis}}$ and use it to obtain predictions on examples in the annotation pool. Based on the ROBERTA$_{\text{analysis}}$ model confidence in the *gold label*, we sample 125 examples from the easy subset, as determined by AFLITE, in a round-robin fashion for each decile between 0 and 1. We repeat this to collect 125 examples from the difficult subset.

Abductive NLI. Since the publicly released α-NLI dataset only contains examples that survived adversarial filtering (‘‘difficult’’ examples), we reached out to the authors to obtain *easy* examples that were filtered out. We source our original examples from the test split of α-NLI. We train a ROBERTA-large model on examples from the train split of α-NLI and follow the same stratified sampling protocol on 125 examples from each of the easy and difficult subsets, according to the confidence of the RoBERTa$_{\text{analysis}}$ in the correct label.

4.2 Paraphrased Example Collection

We obtain 250 examples per dataset (α-NLI, δ-SNLI, δ-ATOMIC, δ-SOCIAL), resulting in 1,000 examples for which we collect paraphrases.

Crowdsourcing. We use Amazon Mechanical Turk to collect paraphrases of the target portions of each example. Workers are shown context sentences and must write a paraphrase of the target sentence(s) according to the definition of label-preserving paraphrasing presented earlier.

We abstract the underlying reasoning task away, presenting α-NLI examples as short stories that require paraphrasing middle sentences, and δ-NLI examples as scenarios with weakening or strengthening evidence. In order to encourage diversity of paraphrases, we display the Jaccard similarity between tokens in the original sentence and the paraphrase as workers typed. Figure 7 shows our annotation interface and instructions for collecting paraphrases of α-NLI and δ-NLI problems respectively.

Workers provide 3 paraphrases of both plausible and implausible hypotheses for α-NLI examples and 3 paraphrases of updates for δ-NLI examples. In the case of α-NLI, we randomly pair together plausible and implausible hypotheses written by the same worker to construct paraphrased examples. Each example was annotated by 3 workers. See Appendix A for more details.

Paraphrase Example Validation. Ensuring semantic equivalence between paraphrased examples and original reasoning problems is essential to our study. Inadvertent removal of the crux of the reasoning problem while paraphrasing may result in invalid examples. Table 3 includes an α-NLI example with paraphrases that were both accepted and rejected from crowdworkers based on our definition of label-preserving paraphrasing. The first and third *accepted* paraphrases both introduce new pieces of information (‘‘mantelpiece’’, ‘‘mounted it on the wall’’) but do not violate the situation evoked by the original problem. In contrast, the first *rejected* paraphrase is incompatible with the situation and the second rejected paraphrase does not retain the plausibility of the hypothesis.

[2] We keep all hyperparameters unchanged from Sakaguchi et al. (2021) with the exception of m, which we reduce from 10,000 to 5,000 to account for the smaller training set partitions, as in Herlihy and Rudinger (2021).

	# original	# paraphrases	mean # paraphrases/ex
α-NLI	250	2098	8.4 ± 1.2
δ-SNLI	250	1980	7.9 ± 1.4
δ-ATOMIC	250	1869	7.5 ± 1.6
δ-SOCIAL	250	1835	7.3 ± 1.8

Table 2: We sample 250 problems from α-NLI and δ-NLI datasets. Total number and mean number of paraphrases depended on validation.

Example	**O1:** George decided to buy a TV. h^+**:** He took it home, unpacked it, and placed it where he wanted it. **O2:** Things had turned out just as he'd hoped.
Accepted	✓ George unboxed the TV and placed it on his mantelpiece. ✓ He removed the plastic and positioned the TV where he had planned to put it. ✓ George brought the new TV home and mounted it on the wall.
Rejected	× George came home and drank a cup of coffee. × He returned the TV.

Table 3: An α-NLI example and paraphrases that were accepted and rejected from crowdworkers.

We opt to have an author validate all paraphrases, each within the context of the problem.[3] A second author and two external annotators annotated a sample of 100 paraphrases, again labeling each as either valid or invalid. We obtain a Fleiss's Kappa (Fleiss, 1971) value of $\kappa = 0.81$ between all validators—the two authors and two external validators. This measures agreement *on the criterion of label-preservation*, reflecting whether a paraphrase written by a crowdworker was of high enough quality to admit into our dataset, as opposed to a measurement of agreement on the correct label given paraphrased examples.

Dataset Overview. Our resulting dataset, ParaNlu, contains 1,000 examples uniformly split across α-NLI, δ-SNLI, δ-ATOMIC, and δ-SOCIAL. Table 2 shows the total number of post-validation paraphrased examples per data split, and the statistics of sizes of buckets.

5 Consistency on Human Paraphrases

We first examine several different models' behavior on ParaNlu to measure robustness to different linguistic expressions. While language models such as RoBERTa are trained on vast amounts of text that may instill some paraphrastic consistency, especially given label-preserving paraphrases that not be semantically equivalent, other non-pretrained models without access to such knowledge may falter. We characterize the progress of models with respect to P_C to understand whether factors such as training setups (training from scratch, supervised finetuning, prompting) and model complexity (ranging from bag-of-words representations to GPT-3) affect consistency.

5.1 Model Variants

We train 5 different types of models per data source. For all models, we use the same set of examples that were used to finetune the RoBERTa$_{\text{analysis}}$ models introduced in §4.1.

Bag of Words. We train bag-of-words models (BoW) using `fasttext`, an off-the-shelf text classification library, with a maximum of 4-grams (Joulin et al., 2017) for 5 epochs with the default learning rate of 0.1.

BiLSTM. We train end-to-end BiLSTM models using the architecture from Conneau et al. (2017) and initialize them with GloVe embeddings (Pennington et al., 2014). We use 3 fully connected layers for classification with max pooling. After tuning on the development sets, models are trained for 10 epochs with early stopping and a batch size of 64.

RoBERTa. We use the RoBERTa$_{\text{analysis}}$ models in §4.1, and add one more setting for defeasible examples in which we finetune a RoBERTa-large model on all combined data across the 3 data sources in δ-NLI, which we refer to as a unified RoBERTa model. All RoBERTa-large models were finetuned for 2 epochs with a learning rate of 2e-5 and a batch size of 32.

DeBERTa. We finetune DeBERTa-v3-large (He et al., 2022) for 2 epochs with a learning rate of 5e-6 and a batch size of 16.

GPT-3. Lastly, we experiment with prompting GPT-3 (Brown et al., 2020) using text-curie-001 (prompts in Table 5). For α-NLI, we randomly sample 36 examples from the training set and include instructions derived from those shown to the crowdworkers that annotated the α-NLI dataset. For δ-NLI, we randomly sample 12 examples per dataset (36 in-context examples total) and include

[3] We explored using NLI models to automatically ensure semantic equivalence, but found it too strict of a formulation to capture the spirit of label-preserving paraphrasing.

	α-NLI						δ-SNLI						δ-ATOMIC						δ-SOCIAL					
	A_{O}	A_{T}	$A_{\text{🧺}}$	P_C	$\widetilde{A}_{\text{🧺}}$	$\widetilde{P}_C$	A_{O}	A_{T}	$A_{\text{🧺}}$	P_C	$\widetilde{A}_{\text{🧺}}$	$\widetilde{P}_C$	A_{O}	A_{T}	$A_{\text{🧺}}$	P_C	$\widetilde{A}_{\text{🧺}}$	$\widetilde{P}_C$	A_{O}	A_{T}	$A_{\text{🧺}}$	P_C	$\widetilde{A}_{\text{🧺}}$	$\widetilde{P}_C$
Lexical (BoW)	44.8	52.4	44.2	100	52.4	100	58.0	55.7	53.7	82.2	52.6	80.9	49.2	53.6	51.4	76.5	53.2	76.5	57.6	61.8	51.4	78.2	54.9	78.5
BiLSTM	53.5	51.6	54.2	100	51.6	100	62.0	68.0	57.6	73.2	60.4	74.2	52.8	67.4	54.2	73.1	61.1	74.8	60.4	72.0	52.4	71.7	59.7	72.8
RoBERTa	53.6	83.5	56.4	69.8	81.5	86.3	51.2	86.7	53.8	74.8	84.6	90.1	53.6	82.6	54.8	76.2	77.9	87.1	51.6	90.9	56.9	74.3	87.8	91.9
DeBERTa-V3	85.6	90.6	73.5	78.4	77.3	79.7	76.8	91.2	70.4	82.8	80.5	84.1	70.4	88.1	66.8	82.7	81.8	87.3	78.4	94.1	71.9	82.2	78.6	83.7
Unified RoBERTa	—	—	—	—	—	—	66.0	85.7	59.9	78.0	81.5	88.6	65.6	84.5	62.8	80.7	78.8	86.8	70.8	90.4	65.7	77.7	83.3	87.7
GPT-3 Curie	46.8	53.5	46.2	80.1	51.4	79.1	52.8	48.9	51.1	89.3	51.3	88	52.0	49.5	52.8	90.2	50.6	89.6	55.6	49.9	57.4	91.3	53.0	90.9

Table 4: Consistency across modeling architectures. A_{O} is accuracy on original examples in ParaNlu, A_{T} is full test set accuracy, $A_{\text{🧺}}$ is accuracy on paraphrased examples ($\widetilde{A}_{\text{🧺}}$ is corrected), and P_C is paraphrastic consistency ($\widetilde{P}_C$ is corrected). No model achieves both high $\widetilde{A}_{\text{🧺}}$ and $\widetilde{P}_C$ (columns highlighted in blue). Optimizing $\widetilde{A}_{\text{🧺}}$ may come at the cost of $\widetilde{P}_C$, or vice versa. We highlight in red the set of *Pareto-optimal* points, or those that are not strictly dominated in $\widetilde{A}_{\text{🧺}}$ or $\widetilde{P}_C$ by any other model.

α-NLI	`Given a story with a beginning and an end, select from two choices the middle sentence that is most plausible and best explains the beginning and end of the story.`
δ-NLI	`Given a Premise sentence, a Hypothesis sentence is defeasible if there exists an Update sentence (consistent with the Premise) such that a human would find the Hypothesis less likely to be true after learning the Update. An Update is called a weakener (abbreviated W) if given a Premise and Hypothesis, a human would most likely find the Hypothesis less likely to be true after learning the Update; if they would find the Hypothesis more likely to be true, then the Update is called a strengthener (abbreviated S). Given a Premise, a Hypothesis, and an Update sentence, assign either W or S to the Update sentence.`

Table 5: Few-shot prompts for α-NLI and δ-NLI tasks. Both prompts were presented to GPT-3 along with 36 in-context examples.

the task definition from Rudinger et al. (2020) in the prompt. Since we cannot reliably extract a softmax distribution over binary classes for our tasks (GPT-3 is not a classification model), we calculate model confidence in a particular class by extracting log probabilities associated with the tokens for both labels and normalize them.

5.2 Results

For all models, we compute P_C (Equation 2). In addition, we undo the biasing effects of the stratified sampling according to model confidence (§4.1) and report a corrected version of paraphrastic consistency, $\widetilde{P}_C$, by weighting the expectations in Equation 2 according to the distribution of model confidences in the correct label of the corresponding test set. We compute four accuracy metrics: (1) accuracy on original examples in the ParaNlu (A_{O}), (2) accuracy on the test set of the original dataset (A_{T}), (3) accuracy on all paraphrases across all buckets ($A_{\text{🧺}}$), and (4) *corrected* accuracy on all paraphrases across all buckets ($\widetilde{A}_{\text{🧺}}$) which is weighted in the same manner as $\widetilde{P}_C$ to undo stratification effects.

Table 4 shows these accuracy metrics along with P_C and $\widetilde{P}_C$ for all models across all data sources. We highlight $\widetilde{A}_{\text{🧺}}$ and $\widetilde{P}_C$, as they are *complementary* metrics meant to jointly assess model performance. Some models optimize $\widetilde{A}_{\text{🧺}}$ at the cost of $\widetilde{P}_C$, or vice versa. To capture models that balance both, we highlight ($\widetilde{A}_{\text{🧺}}$, $\widetilde{P}_C$) points that are *Pareto-optimal* for each dataset. The highest performing model according to $\widetilde{A}_{\text{🧺}}$, RoBERTa, earns a $\widetilde{P}_C$ of around 0.9, indicating room to improve on its paraphrastic consistency **for its accuracy level**. We observe that a GPT-3-curie model with minimal prompt engineering (we simply use the definition of defeasible inference directly from Rudinger et al. (2020)) along with a handful of in-context examples has a $\widetilde{P}_C$ value competitive with a RoBERTa model finetuned on thousands of examples. A stronger GPT-3 variant may better perform defeasible reasoning while maintaining a competitive P_C.

Figure 3 visualizes the relationship between accuracy and $\widetilde{P}_C$ for models on δ-SNLI examples (Figure 8 shows similar plots for δ-ATOMIC and δ-SOCIAL examples). For each model, we plot $\widetilde{A}_{\text{🧺}}$ on the x-axis and $\widetilde{P}_C$ on the y-axis along with two types of supporting curves for the δ-SNLI split of ParaNlu. The curve with the lowest minima (labeled `Min` P_C) indicates the theoretical lower bound for P_C **given a particular accuracy level**: if all the variance in model correctness (Equation 3) is attributable to the paraphrasing variance term present in Equation 7, then the minimum possible value for P_C is $1-2*(Acc*1-Acc)$, where $Acc*(1-Acc)$ is the variance of a Bernoulli random variable with probability Acc. In addition to this theoretical lower bound, we plot curves

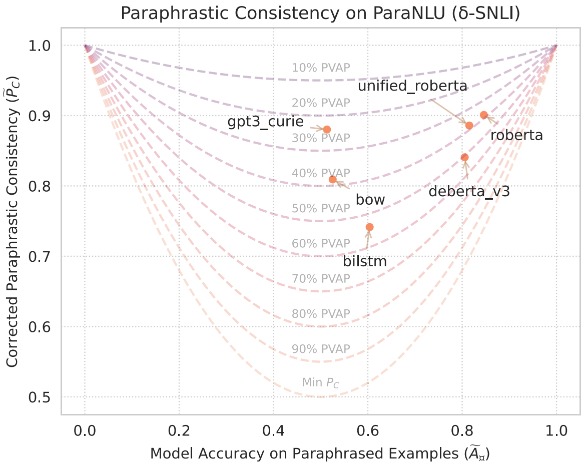


Figure 3: Paraphrastic consistency ($\widetilde{P}_C$) of different models on δ-SNLI paraphrased examples. All models still have room for improvement in $\widetilde{P}_C$ for their accuracy level. Here, we add supporting lines to denote varying levels of the proportion of variance attributable to paraphrasing, or PVAP.

indicating the proportion of total variance attributed to paraphrasing (labeled %PVAP). As P_C increases, less variance is attributed to variance within buckets due to phrasing. Visualizing this relationship makes it clear that accuracy alone provides an incomplete picture of model performance. A perfect model would reside in the top right of Figure 3, not only achieving high accuracy but also high paraphrastic consistency. Models with similar accuracies may have largely different $\widetilde{P}_C$ values, indicating to practitioners how sensitive they are to problem phrasing.

We now turn to a series of experiments to better characterize paraphrases in ParaNlu as well as contributing factors to model's P_C value using our best performing model, RoBERTa.

6 Paraphrase Source

Human-written paraphrases in ParaNlu span all of the transformations delineated by Bhagat and Hovy (2013), sometimes involving more complex reasoning that falls between linguistic and world knowledge. Such paraphrases were elicited by providing humans the entire example, encouraging them to engage with both the reasoning problem itself and the wide scope of possible meaning-preserving transformations. To understand the utility of label-preserving paraphrases, we compare a RoBERTa model's behavior on our human-written paraphrases with paraphrases generated automatically, as previous studies have explored (Verma et al., 2023).

Using our RoBERTa models (§5.1), we probe the relationship between paraphrastic consistency ($\widetilde{P}_C$) and the *source* of paraphrased examples. Are models more robust to the paraphrastic transformations produced by automatic paraphrase generation models, or do they only struggle with the more complex, example-aware transformations made by humans?

Since human paraphrases and model paraphrases are generated by different processes, and are thus drawn from different distributions, they may exhibit different properties. Paraphrase generation models are predisposed to biases arising from n-gram frequency effects.[4] However, reasoning models should exhibit consistency *regardless* of whether correct answers are phrased as high-probability sentences under a language model.

We use two models (§6.2) to automatically paraphrase target sentences in original examples and compare model P_C on automatic and human paraphrases.

6.1 Experimental Setup

For each original example in ParaNlu, we sample paraphrases of targets from generation models. As with humans, we elicit paraphrases of target sentences: update sentences for δ-NLI and both hypothesis sentences for α-NLI. In contrast to the human elicitation process, however, we do *not* provide any context sentences to generation models. While this limits the scope of possible paraphrases, it allows us to gauge the value of exposure to context during paraphrasing.

We adopt a generate-then-validate scheme and have an author again validate all target paraphrases to ensure that their consistency with our definition of label-preserving paraphrasing. In the case of α-NLI examples, where there are multiple target sentences, we randomly pair valid paraphrases together, resampling where necessary when numbers of valid generated hypotheses are unequal.

6.2 Paraphrase Generation Models

Quality-Controlled Paraphrase Generation. We use a QCPG model (Bandel et al., 2022),

[4] For example, a generation model is less likely to paraphrase *''The camera zooms out to show the man spraying the car with soap''* to *''The camera zooms out to servicemen sprinkling the automobile with soap''*.

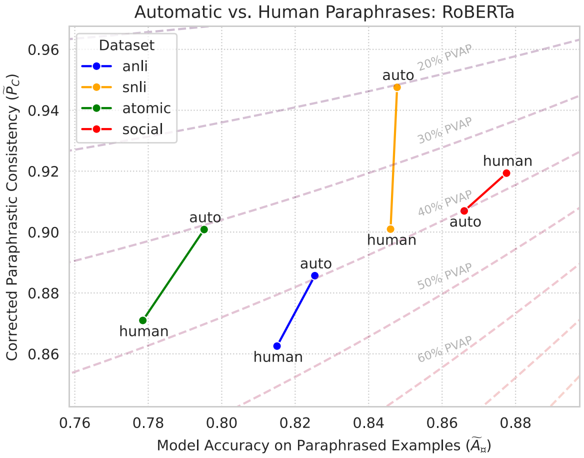


Figure 4: RoBERTa-large model on automatic versus human paraphrases. Models are more consistent on automatic than human paraphrases. Dashed lines indicate varying levels of PVAP.

a controllable paraphrase generation system that conditions on a 3-dimensional vector encoding semantic similarity and lexical and syntactic distances. We pool all paraphrases from a per-example sweep of these hyperparameters.

GPT-3. In addition to a supervised, explicitly controllable paraphrase generation model, we elicit paraphrases from GPT-3 using 10 in-context examples of paraphrases randomly sampled from ParaNlu. Setting temperature to 0.7, we sample 9 paraphrases from text-davinci-002 per target sentence from original examples and similarly validate them to ensure label preservation.

6.3 Results

In total, we generate 7,295 valid paraphrased examples across 1,000 examples by pooling together all *valid* examples from both QCPG and GPT-3 and evaluate our RoBERTa models on these examples. Figure 4 plots $\widetilde{A}_{☗}$ on the x-axis and $\widetilde{P}_C$ on the y-axis for human-generated and automatically-generated paraphrased examples for each dataset. On all datasets with the exception of δ-SOCIAL, we observe that models have a higher $\widetilde{P}_C$ value on automatically generated paraphrased examples than on human-elicited paraphrases. We hypothesize that this pattern may not hold for δ-SOCIAL due to the fact that the dataset does not contain premise sentences, and hence has a smaller scope for more complex transformations involving context. This result suggests that reasoning models may be more robust to the simpler, in-distribution types of paraphrase transformations that automatic paraphrase generation models produce than to those written by human annotators, indicating that over-reliance on evaluation using synthetically generated data may be misleading.

		lex	syn	sem
α-NLI	*automatic*	25.0	20.3	74.3
	human	**35.3**	**26.8**	**64.0**
δ-SNLI	*automatic*	24.1	18.5	76.3
	human	**34.4**	**24.6**	**67.4**
δ-ATOMIC	*automatic*	30.4	19.2	66.7
	human	**36.9**	**22.4**	**58.7**
δ-SOCIAL	*automatic*	30.8	22.2	70.1
	human	**40.0**	**24.5**	**60.1**

Table 6: Lexical (**lex**) and syntactic (**syn**) diversity of human and automatic paraphrases and semantic similarity (**sem**) as compared to original target sentences. Human paraphrases are more diverse than automatic ones: They exhibit higher lexical and syntactic diversity and lower semantic similarity on all datasets (bolded).

Paraphrase Diversity. To dissect this result further, we measure lexical diversity, syntactic diversity, and semantic similarity (Bandel et al., 2022) of target paraphrases and original target sentences. *Lexical distance* is measured by the normalized character-level minimal edit distance between the bag of words (Bandel et al., 2022), and *syntactic distance* is computed as the normalized tree edit distance between the third level constituency parse trees of the original target and the paraphrased target (Iyyer et al., 2018). We measure semantic similarity using BLEURT (Sellam et al., 2020) as in Bandel et al. (2022). Across all four data splits in ParaNlu, human-elicited paraphrases are more lexically and syntactically diverse, as well as less semantically similar to original examples, than automatically generated paraphrases (Table 6). In addition, we find that automatically generated paraphrases are 3–4% more likely to be bidirectionally entailed than human-written paraphrases, as detected by a RoBERTa-large model finetuned on SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), ANLI (Nie et al., 2020), and FEVER (Thorne et al., 2018).

Taken together, these results underscore the benefit of our human annotation to generate ParaNlu—evaluation solely on automatically generated paraphrases, as others have done, is insufficient to fully characterize their robustness.

7 Do Artifacts Explain Inconsistency?

Many NLP datasets are constructed by crowdworkers writing most or all parts of a natural language reasoning problem. While efficient and scalable, this paradigm can give rise to annotation artifacts (Gururangan et al., 2018), or statistical biases in parts of examples that correlate with the correct label (McCoy et al., 2019). For example, Gururangan et al. (2018) found that negation words (*"no"*, *"nothing"*, etc.) in NLI examples are strong indicators of the contradiction label.

One way to detect artifacts in datasets is through a partial-input baseline (Poliak et al., 2018; Feng et al., 2019), a setting in which only the *target* portion of an NLI instance (i.e., the hypothesis in a traditional NLI setup) is used to train a model to predict entailment. When partial-input models achieve high accuracy, it is often indicative of annotation artifacts.

Full-input models that are trained on datasets with annotation artifacts may learn to rely on such shallow signals instead of performing true inferential reasoning. This may lead to lower paraphrastic consistency in the face of alternative phrasings of examples, since they may no longer contain the annotation artifact the model leveraged. Can we attribute a model's issues with paraphrastic consistency on ParaNlu entirely to the presence of annotation artifacts? That is, are models only inconsistent on examples with artifacts, since they are relying on spurious correlations? Or, are they inconsistent even when no artifacts are present?

7.1 Experimental Setup

Rudinger et al. (2020) find that partial-input baselines trained on δ-NLI perform at least 10% better than random chance, indicating the presence of annotation artifacts in their dataset. As such, we focus on δ-NLI for our experiments.[5] For each split of δ-NLI, we train a RoBERTa-large model using only the update sentence as input, keeping the training hyperparameters identical to the full-input RoBERTa models from §5.1. Then, we use these partial-input models to partition buckets in ParaNlu into two subsets: those on which the partial-input model correctly predicted the label from the update sentence of the *original example*, indicating that an artifact is likely present in the original example, and those incorrectly predicted. Using the full-input RoBERTa models from §5.1, we then compute $\widetilde{P}_C$ on both example subsets, and the accuracy of partial-input and full-input models on paraphrased and original examples.

		partial-input			full-input				
	artifacts. . .	A_{O}	$A_{🧺}$	$\widetilde{A}_{🧺}$	A_{O}	$A_{🧺}$	$\widetilde{A}_{🧺}$	P_C	$\widetilde{P}_C$
δ-SNLI	*likely*	100	81.8	54.6	54.3	56.6	85.8	74.7	91.4
	unlikely	0	20.7	6.9	45.5	48.8	79.6	75.1	84.0
δ-ATOMIC	*likely*	100	77.6	53.9	55.6	58.3	78.0	76.4	87.4
	unlikely	0	21.3	8.3	50.9	49.9	79.2	75.9	86.5
δ-SOCIAL	*likely*	100	77.2	58.9	52.9	55.5	85.8	73.9	90.9
	unlikely	0	28.8	8.4	50	58.7	90.7	74.7	93.5

Table 7: Paraphrastic consistency on examples which are likely and unlikely to contain artifacts, as predicted by a partial-input baseline. Inconsistency on both types of examples indicates P_C is attributable to factors *beyond* artifacts.

7.2 Results

Table 7 shows the accuracy metrics for both partial-input and full-input models on original and paraphrased examples in ParaNlu, as well as paraphrastic consistency metrics on both examples that are likely (+) and unlikely (−) to contain artifacts. While not all original examples that a partial-input model predicts correctly *necessarily* have artifacts, we expect that (1) examples with particularly strong artifacts are grouped in the *likely* (+) category, and (2) the *unlikely* (−) category contains a significantly smaller number of examples with strong artifacts. We observe a dramatic drop in the accuracy of a partial-input baseline on original examples (A_{O}) and paraphrased examples ($\widetilde{A}_{🧺}$), indicating that most artifacts detectable with a partial-input model *do not* project through our label-preserving paraphrase process.

Even on examples *unlikely* (−) to contain artifacts, where a full-input model cannot rely on shallow signals, models do still have room to improve their paraphrastic consistency. These results indicate that issues with paraphrastic consistency are attributable to factors beyond the presence of artifacts in examples.

8 Training Dynamics and Paraphrastic Consistency

Lastly, we explore the relationship between different parts of the model training pipeline (e.g., pretraining and finetuning) and paraphrastic consistency. How does consistency *change* as these

[5] We choose δ-NLI instead of α-NLI for this experiment, since the authors of α-NLI released their dataset *after* running adversarial filtering.

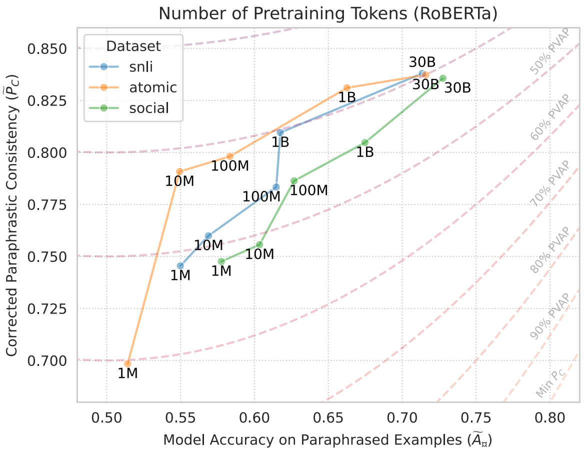


Figure 5: Paraphrastic consistency monotonically increases as a model sees more pretraining tokens, but grows rapidly during early pretraining.

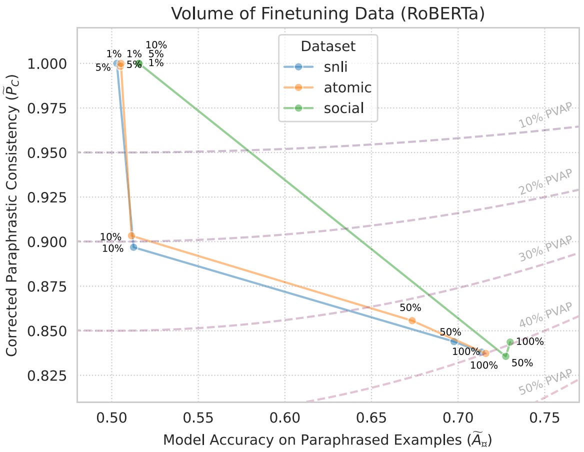


Figure 6: Paraphrastic consistency decreases as a model learns increasingly complex decisions boundaries by seeing more finetuning examples.

training processes progress, and does it change in a similar manner as accuracy? Is it the case that simply increasing the volume of pretraining or finetuning data linearly impacts paraphrastic consistency? We train a series of RoBERTa models and adjust the number of pretraining tokens (§8.1) and finetuning examples (§8.2) to explore how they impact a model's consistency.

8.1 Pretraining and P_C

Experimental Setup. Using the MiniBERTas (Warstadt et al., 2020), a series of RoBERTa models pretrained from scratch on varying numbers of tokens, we compare models trained on 1M, 10M, 100M, and 1B tokens along with RoBERTa-base, which is pretrained on approximately 30B tokens. All models have the same number of parameters as RoBERTa-base (125M), with the exception of the model trained on 1M pretraining tokens which was scaled down in accordance with the smaller volume of pretraining data, and has 45M parameters. We *finetune* all models on the same data (§4.1) and keep all hyperparameters constant (batch size of 64, 2 finetuning epochs, and a learning rate of 5e-6), ensuring direct measurement of the impact of pretraining words on paraphrastic consistency without other confounds.

Results. Figure 5 plots model accuracy on paraphrases ($\widetilde{A}_{¤}$) against paraphrastic consistency ($\widetilde{P}_C$), along with the same supporting curves as in Figure 3 corresponding to decreasing proportions of variance attributable to paraphrasing (PVAP). As expected, pretraining on increasing amounts of data yields both monotonically increasing accuracy and paraphrastic consistency. However, paraphrastic consistency grows more rapidly in the beginning (1M - 100M tokens) as the plots climb steeply between supporting curves and eventually hugs a single PVAP curve past 100M tokens, indicating a slower payoff of more pretraining tokens.

8.2 Finetuning and P_C

After pretraining, models are endowed with the ability to represent natural language inputs but do not know how to perform a particular reasoning task. As such, we expect monotonically increasing accuracy as the model is shown a larger volume of finetuning examples. However, it is unclear how *paraphrastic consistency* changes as the model is exposed to more task-specific examples.

Experimental Setup. For each dataset, we finetune a series of fully pretrained RoBERTa-large models on 1%, 5%, 10%, 50%, and 100% of examples from the training split, sampled at random. δ-ATOMIC has 28.3K training examples, δ-SNLI has 75.2K training examples examples and δ-SOCIAL has 65.3K examples. We sample at the premise-hypothesis level and include all examples that share the same premise and hypothesis to prevent data leakage during evaluation. We hold all training hyperparameters constant, keeping the same configuration as the finetuning in §8.1.

Results. Figure 6 plots corrected paraphrase accuracy ($\widetilde{A}_{¤}$) against corrected paraphrastic consistency ($\widetilde{P_C}$) for models trained on increasing numbers of finetuning examples across all three datasets in δ-NLI. We observe that as the model starts to learn the task at hand and draw increasingly complex decision boundaries, it is more likely to be inconsistent. Models trained on $\leq$5% of the available training examples are highly consistent since they make the same prediction for all examples (thus earning an accuracy of around 50%). As the the model is shown more examples, it makes finer-grained distinctions between examples, in turn impacting its paraphrastic consistency. Though our results show this decrease, it is possible that with even more finetuning data, a model's paraphrastic consistency will start to increase again. This relationship may also be altered if, during finetuning, models are shown increasing amounts of automatically-generated paraphrased examples in order to learn both the task *and* paraphrastic consistency.

9 Related Work

Natural language understanding models may produce different predictions in the face of varying expressions of reasoning problems. A wide range of data generation frameworks have been proposed to study these behaviors in NLP systems. Iyyer et al. (2018) automatically generate paraphrases with specified syntactic templates and measure accuracy on these adversarial examples. Verma et al. (2023) introduce a test set of paraphrases generated with a finetuned T5 model (Raffel et al., 2020) and measure the accuracy of several models. Hu et al. (2019) generate paraphrases of MNLI examples using lexical constraints and evaluate an NLI model on the paraphrased inputs, finding that paraphrasing leads to degraded accuracy. Arakelyan et al. (2024) measure the semantic sensitivity of NLI models by automatically generating examples with FLAN-T5 (Chung et al., 2022) and verifying the generations with bidirectional entailment predicted by pretrained NLI models. While scalable, our findings illustrate that it is insufficient to evaluate models on automatic paraphrases alone, as human-written paraphrases introduce more semantic and pragmatic diversity (Section 6). Moreover, we show that bidirectional entailment as a verification method for generated paraphrases is extremely stringent, precluding us from testing consistency in the face of more challenging label-preserving transformations.

Another body of research studies the creation of adversarial examples to improve model robustness. Nie et al. (2020) construct an NLI benchmark, Adversarial NLI, by developing a model-in-the-loop framework to iteratively produce examples that models cannot correctly solve. Naik et al. (2018) develop a suite of adversarial examples to ''stress test'' common failure modes of NLI models, such as phenomena word overlap or negation. In contrast with these studies, our goal is not to generate a test suite of difficult examples that ''break'' models (Glockner et al., 2018), but rather to carefully measure the role of *paraphrastic* variability in model performance.

Other approaches to measuring robustness also include counterfactual example generation (Srikanth and Rudinger, 2022; Kaushik et al., 2020). Kaushik et al. (2020) recruit humans to create counterfactual examples by minimally editing example text in order to flip the gold label and show that models trained on the original datasets perform poorly on counterfactually-manipulated data. Similarly, Gardner et al. (2020) argue for the creation of evaluative contrast sets, or manual minimal perturbations of dataset examples that change the gold label, in order to probe the decision boundary of models. Our work has a related, but distinct, counterfactual flavor: if an original annotator had chosen to phrase the question differently *with the same target label*, what is the probability that a model's prediction would stay consistent? We aim to estimate, in expectation, the reliability of models when they are faced with different phrasings of the same problem.

Most of these studies measure *accuracy* on adversarial examples as the main determination of robustness. Elazar et al. (2021) instead measure the *consistency* of models with respect to factual knowledge, evaluating whether extracted information from masked language models is invariant to paraphrasing using an agreement-based consistency metric. Our study is similarly concerned with consistency, however we make precise the relationship between accuracy and consistency on natural language reasoning tasks.

10 Conclusion

As more studies investigate the capabilities of LLMs, the ability to disentangle the effects of paraphrastic variability from other target attributes will be an important analytical tool.

This work introduces a new methodology and dataset for measuring paraphrastic consistency, or P_C, of models on natural language reasoning tasks. P_C captures the probability that a model will remain consistent in its prediction given different phrasings of the same underlying reasoning problem. We design P_C as a metric complementary to accuracy, and propose practitioners use it alongside accuracy when diagnosing modeling errors, summarizing a model's performance, or deciding when a model is ready for deployment to users for a particular application.

Our results confirm that paraphrastic sensitivity is present in all models, but decreases with pretraining volume. Because P_C only requires model predictions to be labeled as correct or incorrect, our approach can generalize to any task with binary scoring (and where answers must be invariant to paraphrases). Future work may consider adapting this approach for tasks with more complex or open-ended evaluations.

References

Erik Arakelyan, Zhaoqi Liu, and Isabelle Augenstein. 2024. Semantic sensitivities and inconsistent predictions: Measuring the fragility of NLI models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–444, St. Julian's, Malta. Association for Computational Linguistics.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley Framenet Project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*. https://doi.org/10.3115/980451.980860

Elron Bandel, Ranit Aharonov, Michal Shmueli-Scheuer, Ilya Shnayderman, Noam Slonim, and Liat Ein-Dor. 2022. Quality controlled paraphrase generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 596–609, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.45

Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational Linguistics*, 39(3):463–472. https://doi.org/10.1162/COLI_a_00166

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

Samuel R. 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, Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1075

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, 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 preprint arXiv:2210.11416*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017.

Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1070

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031. https://doi.org/10.1162/tacl_a_00410

Shi Feng, Eric Wallace, and Jordan Boyd-Graber. 2019. Misleading failures of partial-input baselines. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5533–5538, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1554

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378. https://doi.org/10.1037/h0031619

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.48

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. https://doi.org/10.18653/v1/2020.findings-emnlp.117

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague. Association for Computational Linguistics. https://doi.org/10.3115/1654536.1654538

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. https://doi.org/10.18653/v1/P18-2103

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. https://doi.org/10.18653/v1/N18-2017

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2022. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Christine Herlihy and Rachel Rudinger. 2021. MedNLI is not immune: Natural language inference artifacts in the clinical domain. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1020–1027, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-short.129

J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019. Improved lexically constrained decoding for translation and monolingual rewriting. 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 839–850, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1090

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. 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 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1170

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? On the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977. https://doi.org/10.1162/tacl_a_00407

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data.

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. https://doi.org/10.18653/v1/P19-1334

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.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.441

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Charles Sanders Peirce. 1974. *Collected Papers of Charles Sanders Peirce*, volume 5. Harvard University Press.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

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. https://doi.org/10.18653/v1/S18-2023

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1264

Raymond Reiter. 1980. A logic for default reasoning. *Artificial Intelligence*, 13(1-2):81–132. https://doi.org/10.1016/0004-3702(80)90014-4

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, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.418

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106. https://doi.org/10.1145/3474381

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035. https://doi.org/10.1609/aaai.v33i01.33013027

Lenhart Schubert. 2015. Semantic representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29. https://doi.org/10.1609/aaai.v29i1.9759

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.704

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. https://doi.org/10.18653/v1/2022.naacl-main.350

Alessandro Stolfo, Zhijing Jin, Kumar Shridhar, Bernhard Schoelkopf, and Mrinmaya Sachan. 2023. A causal framework to quantify the robustness of mathematical reasoning with language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 545–561, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.32

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A large-scale dataset for fact extraction and VERification. 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 809–819, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1074

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual invariance to spurious correlations in text classification. *Advances in Neural Information Processing Systems*, 34:16196–16208.

Dhruv Verma, Yash Kumar Lal, Shreyashee Sinha, Benjamin Van Durme, and Adam Poliak. 2023. Evaluating paraphrastic robustness in textual entailment models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 880–892, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-short.76

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.emnlp-main.16`

N. A. Weiss, P. T. Holmes, and M. Hardy. 2006. *A Course in Probability*. Pearson Addison Wesley.

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. `https://doi.org/10.18653/v1/N18-1101`

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics. `https://doi.org/10.18653/v1/D18-1009`

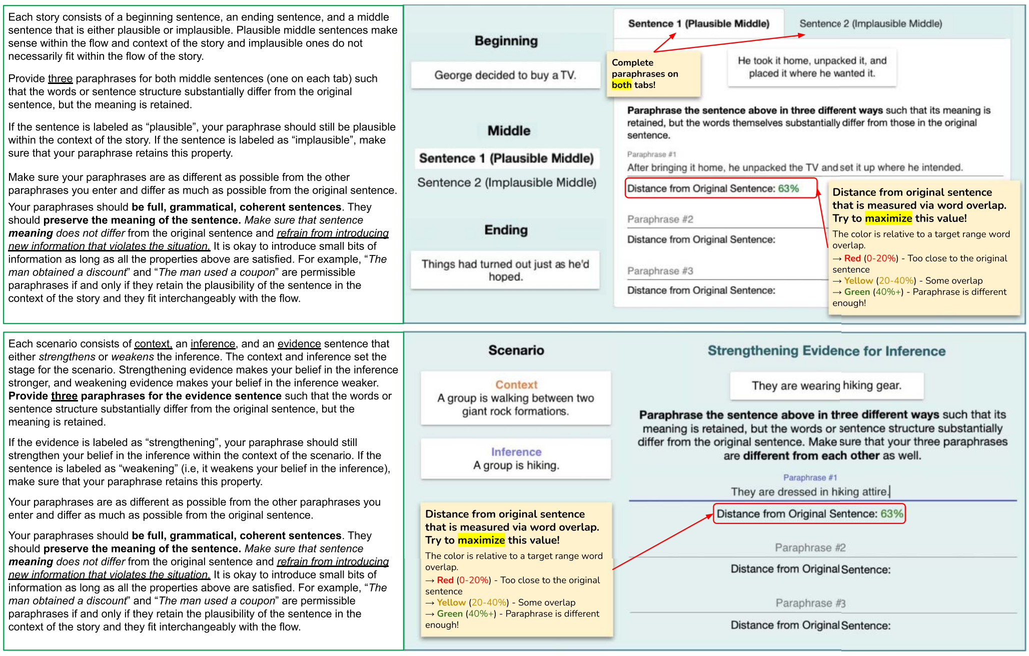
A Crowdsourcing ParaNlu

We collect paraphrases in ParaNlu using Amazon Mechanical Turk. The instructions and annotation interface shown to crowdworkers is shown in Figure 7.

Workers provide 3 paraphrases of both plausible and implausible hypotheses for α-NLI examples and 3 paraphrases of updates for δ-NLI examples. We include a distance widget in our interface that computes the Jaccard similarity between the entered paraphrase and the original text to encourage lexical diversity. Each example was annotated by 3 workers. Workers were paid US$12/hour on average and were required to be native English speakers with a 95% or more HIT acceptance rate on at least 100 HITs.

B Paraphrastic Consistency

Figure 8 shows model accuracy plotted against corrected paraphrastic consistency of all models tested for the δ-ATOMIC and δ-SOCIAL splits of ParaNlu.



Each story consists of a beginning sentence, an ending sentence, and a middle sentence that is either plausible or implausible. Plausible middle sentences make sense within the flow and context of the story and implausible ones do not necessarily fit within the flow of the story.

Provide three paraphrases for both middle sentences (one on each tab) such that the words or sentence structure substantially differ from the original sentence, but the meaning is retained.

If the sentence is labeled as “plausible”, your paraphrase should still be plausible within the context of the story. If the sentence is labeled as “implausible”, make sure that your paraphrase retains this property.

Make sure your paraphrases are as different as possible from the other paraphrases you enter and differ as much as possible from the original sentence.

Your paraphrases should **be full, grammatical, coherent sentences**. They should **preserve the meaning of the sentence.** *Make sure that sentence **meaning** does not differ* from the original sentence and *refrain from introducing new information that violates the situation.* It is okay to introduce small bits of information as long as all the properties above are satisfied. For example, “*The man obtained a discount*” and “*The man used a coupon*” are permissible paraphrases if and only if they retain the plausibility of the sentence in the context of the story and they fit interchangeably with the flow.

Beginning

George decided to buy a TV.

Middle

Sentence 1 (Plausible Middle)

Sentence 2 (Implausible Middle)

Ending

Things had turned out just as he'd hoped.

Sentence 1 (Plausible Middle) Sentence 2 (Implausible Middle)

Complete paraphrases on **both** tabs!

He took it home, unpacked it, and placed it where he wanted it.

Paraphrase the sentence above in three different ways such that its meaning is retained, but the words themselves substantially differ from those in the original sentence.

Paraphrase #1

After bringing it home, he unpacked the TV and set it up where he intended.

Distance from Original Sentence: 63%

Paraphrase #2

Distance from Original Sentence:

Paraphrase #3

Distance from Original Sentence:

Distance from original sentence that is measured via word overlap. Try to maximize this value!

The color is relative to a target range word overlap.

→ **Red** (0-20%) - Too close to the original sentence

→ **Yellow** (20-40%) - Some overlap

→ **Green** (40%+) - Paraphrase is different enough!

Each scenario consists of context, an inference, and an evidence sentence that either *strengthens* or *weakens* the inference. The context and inference set the stage for the scenario. Strengthening evidence makes your belief in the inference stronger, and weakening evidence makes your belief in the inference weaker. **Provide three paraphrases for the evidence sentence** such that the words or sentence structure substantially differ from the original sentence, but the meaning is retained.

If the evidence is labeled as “strengthening”, your paraphrase should still strengthen your belief in the inference within the context of the scenario. If the sentence is labeled as “weakening” (i.e, it weakens your belief in the inference), make sure that your paraphrase retains this property.

Your paraphrases are as different as possible from the other paraphrases you enter and differ as much as possible from the original sentence.

Your paraphrases should **be full, grammatical, coherent sentences**. They should **preserve the meaning of the sentence.** *Make sure that sentence **meaning** does not differ* from the original sentence and *refrain from introducing new information that violates the situation.* It is okay to introduce small bits of information as long as all the properties above are satisfied. For example, “*The man obtained a discount*” and “*The man used a coupon*” are permissible paraphrases if and only if they retain the plausibility of the sentence in the context of the story and they fit interchangeably with the flow.

Scenario

Context
A group is walking between two giant rock formations.

Inference
A group is hiking.

Distance from original sentence that is measured via word overlap. Try to maximize this value!

The color is relative to a target range word overlap.

→ **Red** (0-20%) - Too close to the original sentence

→ **Yellow** (20-40%) - Some overlap

→ **Green** (40%+) - Paraphrase is different enough!

Strengthening Evidence for Inference

They are wearing hiking gear.

Paraphrase the sentence above in three different ways such that its meaning is retained, but the words or sentence structure substantially differ from the original sentence. Make sure that your three paraphrases are **different from each other** as well.

Paraphrase #1

They are dressed in hiking attire.

Distance from Original Sentence: 63%

Paraphrase #2

Distance from Original Sentence:

Paraphrase #3

Distance from Original Sentence:

Figure 7: Annotation instructions and interface for collecting paraphrases of α-NLI (top) and δ-NLI (bottom) reasoning problems. Workers must write three label-preserving paraphrases.

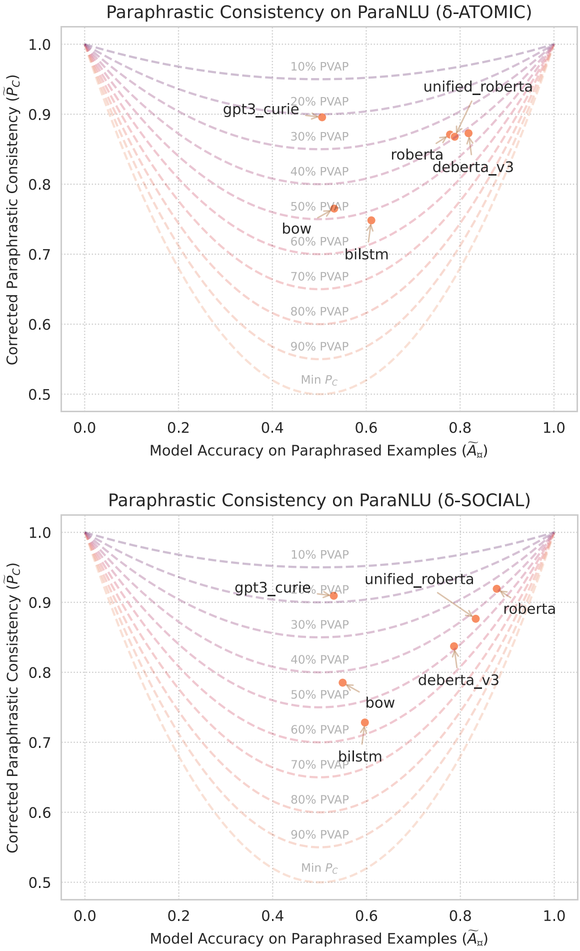


Figure 8: Paraphrastic consistency ($\widetilde{P}_C$) of different models on δ-ATOMIC and δ-SOCIAL paraphrased examples.