With a Little Push, NLI Models *can* Robustly and Efficiently Predict Faithfulness

Julius Steen Juri Opitz Anette Frank Katja Markert
Department of Computational Linguistics
Heidelberg University
69120 Heidelberg, Germany
(steen|opitz|frank|markert)@cl.uni-heidelberg.de

Abstract

Conditional language models still generate unfaithful output that is not supported by their input. These unfaithful generations jeopardize trust in real-world applications such as summarization or human-machine interaction, motivating a need for automatic faithfulness metrics. To implement such metrics, NLI models seem attractive, since they solve a strongly related task that comes with a wealth of prior research and data. But recent research suggests that NLI models require costly additional machinery to perform reliably across datasets, e.g., by running inference on a cartesian product of input and generated sentences, or supporting them with a question-generation/answering step.

In this work we show that pure NLI models *can* outperform more complex metrics when combining task-adaptive data augmentation with robust inference procedures. We propose: (1) Augmenting NLI training data to adapt NL inferences to the specificities of faithfulness prediction in dialogue; (2) Making use of both entailment and contradiction probabilities in NLI, and (3) Using Monte-Carlo dropout during inference. Applied to the TRUE benchmark, which combines faithfulness datasets across diverse domains and tasks, our approach strongly improves a vanilla NLI model and significantly outperforms previous work, while showing favourable computational cost.

1 Introduction

Conditional language models suffer from a tendency to *hallucinate* information (Maynez et al., 2020), resulting in generations that are not faithful to their input documents, which limits the trustworthiness of such models. This raises a need for automatic faithfulness metrics. In this context, models trained on natural language inference (NLI) (Bowman et al., 2015) are attractive since, intuitively, a generation being *faithful* implies it must be *entailed* by the source (Falke et al., 2019).

However, pure NLI models have seen mixed success in faithfulness evaluation (Falke et al., 2019; Kryscinski et al., 2020; Wang et al., 2020; Maynez et al., 2020). While in recent evaluation on the TRUE benchmark (Honovich et al., 2022), which contains datasets from knowledge-grounded dialogue, summarization and paraphrasing, NLI-derived metrics perform best overall, they require impractically large models, or costly additional machinery such as question generation and answering models at inference, while still showing robustness issues. Thus we ask: *What is still needed for pure NLI models to perform robustly across faithfulness datasets – while remaining cheap enough to serve as a lean and practical evaluation tool?*

We enhance a relatively small NLI model to make it work robustly across tasks in three ways:

Task-Adaptive Data Augmentation. In NLI, a hypothesis must be fully entailed by its supporting premise. However, in faithfulness, not all parts of the generation always need to be grounded. We identify an instance of this phenomenon in dialogue where parts of a turn can fulfill communicative functions such as hedging or establishing emotional connection and are often disregarded in faithfulness annotation. Hence, when applying NLI models to *complete dialogue turns* that may include statements irrelevant for grounding, we run a risk of producing incorrect unfaithfulness predictions.

To alleviate this issue, we propose a simple **data augmentation** method to adapt NLI models to genres where they need to be aware of statements that must be exempt from NLI-based faithfulness evaluation. Our approach is computationally attractive, as it avoids an increase of cost at inference time.

Integration of NLI Contradiction Scores. Existing NLI faithfulness metrics typically use the entailment score for their predictions (Honovich et al., 2022; Falke et al., 2019; Kryscinski et al., 2020). However, Chen and Eger (2022) show that subtracting the contradiction score from the entail-

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 914–924

ment score (referred to as e-c) can improve NLI performance in certain evaluation tasks. We show that there also is a strong positive effect of e-c for faithfulness prediction, and demonstrate that this is due to a high contradiction probability being a more reliable predictor of unfaithfulness than low entailment probability.

Monte-Carlo Dropout Inference. Applying NLI models to faithfulness prediction involves a domain shift from largely human-written data to automatically generated text. To make NLI model scores more robust under this shift, we propose to use Monte-Carlo dropout during inference (Srivastava et al., 2014). This essentially creates a cheap *ensemble* and has been shown to deal better with noisy labels (Goel and Chen, 2021). This approach leads to consistent score improvements in our tasks.

The combination of all modifications not only strongly improves over a baseline NLI model, but also outperforms all other metrics on TRUE, on average, while being **cheaper** and **smaller**.[1]

2 Method Details

2.1 Task-adaptive Data Augmentation

To illustrate that task requirements can be incompatible between faithfulness and NLI, consider the following instance from the Q2 dialogue corpus (Honovich et al., 2021) that is labelled as faithful:

> **Grounding:** American pancakes are similar to Scotch pancakes or drop scones.
> **Generation:** yes , i love american pancakes , they are like scotch pancakes

From an NLI perspective, the generation is clearly not entailed, since the statement "I love american pancakes" is not supported by the input.

To better prepare an NLI system for such genre or task-specific cases, we manually curate a small list of statements that should not influence the faithfulness prediction. We augment NLI data from the ANLI corpus (Nie et al., 2020) by adding a randomly chosen phrase from this set to each instance, while preserving the label. We then train an already fine-tuned NLI model on a concatenation of these augmented samples and original ANLI data. For training details see Appendix A.

[1]All code is available at https://github.com/julmaxi/with_a_little_push

2.2 Monte-Carlo Dropout

To compute scores under Monte-Carlo dropout, we randomly sample k dropout masks and compute the average of the model predictions. We set $k = 15$, since preliminary experiments showed that performance did not profit from additional samples.

3 Experimental Setup

We run experiments on TRUE (Honovich et al., 2022), a benchmark that compiles a wide variety of faithfulness tasks in a standardized format. It contains summarization (Pagnoni et al., 2021; Maynez et al., 2020; Wang et al., 2020; Fabbri et al., 2021), knowledge-grounded dialog (Honovich et al., 2021; Gupta et al., 2022; Dziri et al., 2022)[2] and paraphrasing (Zhang et al., 2019) datasets.[3] Following recommendations in TRUE, we evaluate using Area under the ROC Curve (AUC).

As our `BASE` model, we use the DeBERTa-large (He et al., 2020) model of Laurer et al. (2022), trained on MultiNLI (Williams et al., 2018), Fever-NLI (Thorne et al., 2018), ANLI (Nie et al., 2020), LingNLI (Parrish et al., 2021) and WANLI (Liu et al., 2022). The metric `All` uses all three of our proposed modifications to `Base`. We also investigate a variant without MC dropout inference (`-MC`) as a more cost efficient alternative.

We compare to the strongest models on TRUE:

`T5 ANLI` (Honovich et al., 2022) is a T5-11B (Raffel et al., 2020) model trained on ANLI.[4]

`SummacZS` (Laban et al., 2022) evaluates an NLI model on all pairs of input and generated sentences and then averages maximum entailment probabilities for each generated sentence.

`Q2` (Honovich et al., 2021) combines a question generation/answering pipeline with an NLI score.

Finally, Honovich et al. (2022) introduce a strong ensemble of these 3 methods (`Eorig`). To further verify our approach, we construct a new ensemble (`Eour`) by replacing `T5` with `All`.

4 Results

Table 1 shows the AUC scores for each metric. Our model `All` not only significantly improves over

[2]TRUE uses an earlier variant of BEGIN that is described in https://arxiv.org/pdf/2105.00071v1.pdf

[3]TRUE also has a fact-checking part, which was not included in average metric performance. We also exclude it here, as our base NLI model was trained on parts of it.

[4]The original `T5` model is also pretrained on GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) data, which contains additional NLI data.

Method	Q2	SummacZS	T5 ANLI	Base	-MC	All	Eorig	Eour
Summarization								
Frank	$_{85.4}87.8_{90.0}$	$_{86.7}89.1_{91.1}$	$_{87.3}\mathbf{89.4}_{91.2}$	$_{83.1}85.6_{88.0}$	$_{84.2}86.6^{\dagger}_{88.9}$	$_{85.5}87.7^{\dagger}_{89.8}$	$_{89.4}91.2_{93.0}$	$_{89.7}91.5_{93.2}$
MNBM	$_{65.6}68.7_{71.7}$	$_{68.6}71.3_{74.1}$	$_{75.5}\mathbf{77.9}_{80.2}$	$_{71.7}74.6_{77.4}$	$_{70.1}73.5_{76.6}$	$_{71.3}74.5_{77.4}$	$_{74.0}76.6_{79.4}$	$_{73.6}76.4_{79.2}$
SummEval	$_{75.9}78.8_{81.4}$	$_{79.4}\mathbf{81.7}_{83.9}$	$_{78.0}80.5_{83.0}$	$_{69.6}72.8_{75.8}$	$_{72.3}75.2^{\dagger}_{78.1}$	$_{73.2}76.1^{\dagger}_{78.8}$	$_{80.4}82.9_{85.4}$	$_{80.3}83.0_{85.3}$
QAGS-X	$_{65.5}70.9_{76.2}$	$_{73.1}78.1_{82.9}$	$_{79.5}\mathbf{83.8}_{88.2}$	$_{76.9}81.6_{86.5}$	$_{77.7}82.2_{86.8}$	$_{76.3}81.1_{85.4}$	$_{80.4}84.8_{88.9}$	$_{79.4}83.8_{88.0}$
QAGS-C	$_{79.1}\mathbf{83.5}_{87.9}$	$_{76.3}80.9_{85.2}$	$_{77.5}82.1_{86.7}$	$_{68.7}74.1_{79.3}$	$_{73.0}78.4^{\dagger}_{82.9}$	$_{73.2}78.0^{\dagger}_{82.9}$	$_{83.5}87.7_{91.3}$	$_{83.1}86.7_{90.3}$
Dialogue								
BEGIN	$_{77.2}79.7_{82.2}$	$_{79.2}82.0_{84.6}$	$_{80.3}\mathbf{82.6}_{85.1}$	$_{77.5}80.4_{82.9}$	$_{75.7}78.5_{81.4}$	$_{76.4}79.3_{82.3}$	$_{84.1}86.2_{88.2}$	$_{82.1}84.7_{87.1}$
DialFact	$_{85.4}86.1_{86.8}$	$_{83.3}84.1_{84.8}$	$_{76.8}77.7_{78.6}$	$_{81.0}81.8^{*}_{82.5}$	$_{91.3}91.8^{*\dagger x}_{92.3}$	$_{92.0}\mathbf{92.5}^{*\dagger x}_{93.0}$	$_{89.9}90.4_{91.0}$	$_{94.1}94.5^{x}_{94.9}$
Q2	$_{78.8}80.9_{83.0}$	$_{74.9}77.4_{79.7}$	$_{70.3}72.7_{75.2}$	$_{77.5}79.8^{*}_{82.0}$	$_{87.2}88.8^{*\dagger x}_{90.3}$	$_{87.8}\mathbf{89.4}^{*\dagger x}_{90.9}$	$_{80.8}82.8_{84.9}$	$_{86.8}88.5^{x}_{90.1}$
Paraphrasing								
PAWS	$_{89.1}89.7_{90.3}$	$_{87.5}88.2_{88.7}$	$_{85.7}86.4_{87.1}$	$_{87.2}87.8^{*}_{88.4}$	$_{88.4}89.0^{*\dagger}_{89.6}$	$_{89.4}\mathbf{90.0}^{*\dagger}_{90.5}$	$_{90.7}91.2_{91.7}$	$_{91.8}92.3^{x}_{92.8}$
Avg	$_{79.7}80.7_{81.7}$	$_{80.4}81.4_{82.3}$	$_{80.6}81.5_{82.4}$	$_{78.8}79.8_{80.8}$	$_{81.7}82.7^{\dagger}_{83.6}$	$_{82.2}\mathbf{83.2}^{*\dagger}_{84.1}$	$_{85.1}86.0_{86.8}$	$_{86.0}86.8^{x}_{87.7}$

Table 1: AUC scores for all models on TRUE. Small numbers indicate 95% CIs computed via bootstrap. $*$ indicates statistically significant improvement over T5; †: statistically sign. improvement over Base; ^{x}: statistically sign. improvement over Eorig ($p \leq 0.05$, approximate randomization test). Best non-ensemble models in bold.

Base on six out of nine corpora, but also significantly outperforms all other competitors on average, while being more computationally efficient.

As expected, we find the biggest gains in dialogue, where the All model even outperforms Eorig on 2 out of 3 corpora. We do not improve on BEGIN, which is likely due to bias in the dataset construction, which we elaborate on in Section 5.1. On the summarization part, All improves significantly over Base on 3 out of 5 corpora, while not significantly harming performance on any corpus. However, it still falls short of the best models in TRUE. The strong showing of T5 on these corpora suggests that this might be alleviated with a stronger base model.

Overall, a very similar behaviour is exhibited by -MC, presenting an attractive option when the added overhead of multiple samples is undesirable.

Eour is on par with Eorig, despite massively reduced costs; it even significantly outperforms it on two dialog and the paraphrasing corpora.

We also investigate the performance of each individual modification to our model (Table 2). They all improve average scores, while only leading to a notable decrease on BEGIN for both *e-c* and dialogue augmentations and on MNBM for *e-c* .

Outside of dialogue, we find that the augmentation methods have a positive impact on PAWS, as well as all summarization corpora that are at least partially based on summaries for the CNN/DM dataset (Hermann et al., 2015) (Frank, QAGS-C, and SummEval). While we do not have a definitive explanation for this phenomenon, we hypothesize that on these datasets our augmentations aid in making the model robust in the presence of noise or irrelevant context since our augmentations are label-neutral and must similarly be 'ignored' during training.

Corpus	+*e-c*	+MC	+Aug.
Frank	$_{-0.0}+0.3_{+0.5}$	$_{+0.1}+0.9_{+1.8}$	$_{+0.3}+1.0_{+1.7}$
MNBM	$_{-2.1}-0.8_{+0.5}$	$_{+1.4}+2.1_{+2.9}$	$_{-0.4}+0.0_{+0.6}$
SummEval	$_{+0.7}+1.0_{+1.3}$	$_{+0.1}+1.2_{+2.3}$	$_{+0.6}+1.6_{+2.6}$
QAGS-X	$_{-0.4}+0.3_{+0.9}$	$_{-1.5}-0.2_{+1.1}$	$_{-0.3}+0.9_{+2.1}$
QAGS-C	$_{+0.5}+1.2_{+2.0}$	$_{-1.6}-0.1_{+1.5}$	$_{+2.2}+3.5_{+5.0}$
BEGIN	$_{-3.0}-1.1_{+0.6}$	$_{+0.0}+0.6_{+1.3}$	$_{-1.6}-1.0_{-0.5}$
DialFact	$_{+8.3}+9.1_{+9.9}$	$_{+1.1}+1.3_{+1.5}$	$_{+3.1}+3.3_{+3.5}$
Q2	$_{+5.1}+6.5_{+7.9}$	$_{-0.4}-0.0_{+0.4}$	$_{+3.5}+4.2_{+5.0}$
PAWS	$_{+0.3}+0.4_{+0.5}$	$_{+1.1}+1.3_{+1.4}$	$_{+0.8}+0.9_{+1.0}$
Avg	$_{+1.6}+1.9_{+2.2}$	$_{+0.5}+0.8_{+1.1}$	$_{+1.4}+1.6_{+1.9}$

Table 2: AUC differences for individual modifications of Base. Small numbers: 95% CIs (bootstrap resampling).

5 Analysis

5.1 Effect of Dialogue Adaptation

We investigate whether the improvements via our augmentation approach are indeed due to them improving the handling of personal statements.

We use the occurrences of the pronoun *I* in a generation as a proxy measure[5] and compute its correlation with human labels and metrics (see Table 3). On both Q2 and Dialfact, our proxy measure, while uncorrelated with human labels, is strongly correlated with the scores of both Base and T5. This indicates these metrics indeed tend to incorrectly reject generations with personal statements. All on the other hand reduces this dependency.

Our results also help explain why All fails to improve on BEGIN, since BEGIN gold labels are

[5]We use spacy (spacy.io) for POS tagging to identify pronouns.

Method	(BEGIN)	Q2	DialFact
T5	(-0.27)	-0.40	-0.13
Base	(-0.28)	-0.32	-0.10
All	(-0.19)	-0.19	0.04
Gold Label	(-0.35)	-0.03	0.05

Table 3: Kendall's τ correlations of gold labels/system scores with first person pronoun occurrence. BEGIN shows a strong negative correlation which we attribute to model-induced dataset bias (see Appendix B).

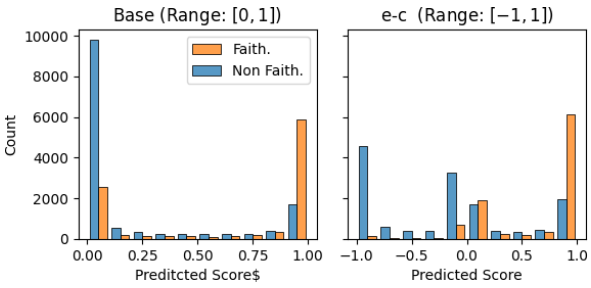

Figure 1: Histogram of the score distributions with and without e-c for faithful and non-faithful instances.

negatively correlated with first person pronouns. This is likely due to a bias in dataset construction: The BEGIN dataset used in TRUE has generations from two models, one of which is both more likely to generate pronouns and more likely to generate unfaithful output (see Appendix B).

5.2 Effect of integrating contradiction scores

To isolate the effect of e-c we compare score distributions of Base and Base+e-c in Figure 1. The left-hand side of the figure shows that in Base ca. 2700 faithful instances are predicted as non-entailed (i.e., e-score near 0), which implies they are labelled as contradictory or neutral. e-c , on the other hand, further differentiates these instances into instances with high contradiction (negative e-c score) and high neutral probability (e-c score near 0). We observe that almost all low-scoring faithful generations are classified as neutral, whereas nearly all instances that are classified as contradictory are indeed unfaithful. Where Base has no way to make use of this information, e-c allows to reliably label contradictory instances as unfaithful.

5.3 Cost comparison to other approaches

There is increasing awareness of the resource-hungriness of deep learning (Strubell et al., 2019). Especially for faithfulness, cheap and reliable metrics are critical, given rising demands for NLG in research and industry. Table 4 shows that our model

Method	AUC↑	Param·10^6↓	Model calls↓
SummacZS	80.7	355	#snt×#snt
T5 ANLI	81.5	11,000	1
Q2	81.4	220 + 355 + 355	#Q × (Ql + 2)
-MC	82.7	350	1
All	83.2	350	15

Table 4: Performance vs. cost analysis

	w/ Five Augmentations				No Aug.
Dataset	Avg.	Std.	Min	Max	Avg.
Frank	$86.7_{-1.0}$	0.4	85.8	87.6	86.2
MBNM	$74.4_{-0.1}$	0.4	73.7	74.9	75.1
SummEval	$75.2_{-0.9}$	0.5	74.5	76.0	74.3
QAGS-X	$81.6_{+0.5}$	0.5	80.8	82.4	80.7
QAGS-C	$76.4_{-1.6}$	0.8	74.7	77.9	75.2
DialFact	$92.1_{-0.4}$	0.2	91.5	92.3	91.2
BEGIN	$79.6_{+0.3}$	0.5	79.0	80.6	80.9
Q2	$88.8_{-0.6}$	0.3	88.1	89.2	86.3
PAWS	$89.7_{-0.3}$	0.1	89.5	90.0	89.3
Avg.	$82.7_{-0.5}$	0.2	82.3	82.9	82.1

Table 5: Results of our phrase selection robustness analysis. For each run, we sample five phrases, recreated our dataset and retrain our model. We repeat this process ten times and report the average, as well as the standard deviation, minimum and maximum scores of the runs. Small numbers indicate difference to the original scores. All results were computed using e-c and MC dropout. For better comparison, we also report the scores of a model without any augmentation (i.e. without any additional training) with e-c and MC dropout.

requires fewer parameters than any other metric, including a more than 30x reduction compared to T5. During inference our model always requires a constant number of calls which can be reduced to a single call when ablating MC dropout. On the other hand, the number of calls in SummacZS scales with the number of input and output sentences. Q2 needs to generate questions by calling an auto-regressive QG model n times, where n factors in the amount and length of questions (#Q×Ql), answer #Q questions with the QA model and finally check #Q answers with an NLI model (#Q × 2).

In sum, our model compares favourably with other approaches, while also allowing for a performance/cost tradeoff by forgoing MC dropout.

5.4 Phrase Selection Robustness

To ensure that our augmentation is robust and not overly reliant on any particular choice of phrases, we repeat our dataset augmentation process multiple times with five randomly chosen augmentation phrases out of the original ten. We sample ten such datasets and retrain our model for each. Table 5 shows the average score, minimum and maxi-

mum score, as well as the standard deviation of the scores. We also report results of a model with both MC dropout and e-c but without any additional training and augmentations to directly quantify whether the augmentations are still helpful in their reduced form. This corresponds to applying MC dropout and e-c to `Base`.

As expected, we find that reducing the variety of available phrases leads to a drop in performance across almost all datasets, compared to `All`. The only exception is BEGIN, where we instead see a slight improvement. This is likely to be related to the construction of BEGIN (see the discussion in Section 5.1).

When comparing our limited augmentation models to the non-augmented model, we find that they still outperform the non-augmented model in almost all cases. In particular for Q2 and DialFact, for which we expect the strongest impact of our augmentations, we find that even the worst run still outperforms non-augmented model. This suggests that our augmentations can robustly adapt the model to the dialogue task.

Finally, we observe a relatively large drop in scores for all datasets that are at (least partially) derived from CNN/DM (Frank, SummEval and QAGS-C). This mirrors our earlier observation in Section 4 that these datasets profit from our augmentation procedure.

6 Related Work

Previous work on the utility of NLI for faithfulness led to mixed conclusions. In summarization, Falke et al. (2019) and Kryscinski et al. (2020) find out-of-the-box models have only limited utility in a faithfulness setting. In Wang et al. (2020), an NLI model is outperformed by a question generation/answering (QA/QG)-based method. In contrast, Maynez et al. (2020) find that a similar NLI model vastly outperforms a QA/QG metric on their data. In knowledge-grounded dialogue, Dziri et al. (2022), Gupta et al. (2022) and Honovich et al. (2021) find out-of-the-box models underperform.

To improve NLI models for faithfulness in summarization, Kryscinski et al. (2020) propose FactCC, which is trained on artificially noised summaries. Utama et al. (2022) propose a controllable generation model to generate artificial faithfulness data. In knowledge-grounded dialogue, Dziri et al. (2022) and Gupta et al. (2022) combine noising techniques to generate additional training data for NLI-based faithfulness models. In contrast to our work, these approaches a) generate training data from external sources, instead of directly augmenting NLI data, and b) do not explicitly focus on reconciling differences between NLI and faithfulness with their augmentation. Outside of augmentation-based approaches, Goyal and Durrett (2020) propose to train NLI models to label faithfulness at the dependency arc level.

7 Conclusion

We have demonstrated that with a small number of focused adaptations, even a relatively small NLI model can robustly predict faithfulness. We have:

1. Shown that NLI-based metrics can be incompatible with task-specific requirements and identified and fixed one such incompatibility in dialogue with an augmentation strategy.

2. Demonstrated the importance of contradiction probability for scoring and that the underlying mechanism is the high reliability of NLI contradiction scores for detecting unfaithfulness

3. Shown that using Monte-Carlo dropout improves metric performance.

Our improved NLI model significantly improves over its baseline across many corpora and outperforms all competitors in average score on TRUE, while being much more efficient at inference.

Our work suggests that strong improvements are possible for NLI-based faithfulness metrics, by combining data augmentation with adapted NLI score computation. We hope this finding will spurn advances in cheap and robust NLI for faithfulness.

8 Limitations

Some of the summarization datasets annotated for faithfulness are relatively small, which makes score estimates uncertain. Furthermore, many datasets contain only output from a limited number of generation systems, which makes it hard to properly account for potential biases towards certain generation systems that may confound scores (see Pagnoni et al. (2021)). These concerns are, however, alleviated to some extent since we study trends across many independently created datasets, which makes it less likely for a single bias to persist in all of them. Furthermore the availability of generation and thus annotated faithfulness data limits our experiments to English. Finally, it remains

unclear whether our results would still provide advantages when applied to larger models such as T5-11B, whose parameter count makes experimentation infeasible on the hardware available to us.

9 Ethics Statement

Faithfulness metrics help reduce the amount of incorrect information generated by NLG systems, reducing the risk associated which such generations. However, faulty or unreliable faithfulness metrics might cause harm by incorrectly classifying faithful content as unfaithful and vice versa.

We run all experiments on publicly available data that has been specifically constructed for faithfulness evaluation. The underlying publication has been published at a conference whose review process involved an ethics review. For a specific discussion of the human effort involved in creation of the datasets we refer the reader to the original publications.

References

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.

Yanran Chen and Steffen Eger. 2022. Menli: Robust evaluation metrics from natural language inference. *arXiv preprint arXiv:2208.07316*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022. Evaluating Attribution in Dialogue Systems: The BEGIN Benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083. **Note:** TRUE uses an earlier version of the BEGIN dataset. The version used in TRUE is described in an earlier preprint at `https://arxiv.org/pdf/2105.00071v1.pdf`.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Purvi Goel and Li Chen. 2021. On the robustness of monte carlo dropout trained with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2219–2228.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. DialFact: A benchmark for fact-checking in dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801, Dublin, Ireland. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1693–1701, Cambridge, MA, USA. MIT Press.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Moritz Laurer, W v Atteveldt, Andreu Casas, and Kasper Welbers. 2022. Less annotating, more classifying–addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli.

Alisa Liu, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2022. Wanli: Worker and ai collaboration for natural language inference dataset creation. *arXiv preprint arXiv:2201.05955*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. 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.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alexia Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R. Bowman. 2021. Does putting a linguist in the loop improve NLU data collection? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4886–4901, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. 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 704–718, Online. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

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.

Prasetya Utama, Joshua Bambrick, Nafise Moosavi, and Iryna Gurevych. 2022. Falsesum: Generating document-level NLI examples for recognizing factual inconsistency in summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2763–2776, Seattle, United States. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

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*

Introductory Statements
Here is what I know:
yep. Also
Sure! Here is what I know:
Hedging
I am not sure, but
I am not sure but I do know that
I do not have information on this but
I think
I believe
Sentiment
I love that!
I like that!

Table 6: Manually curated list of dialogue phrases

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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

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 Augmentation Training Details

A.1 Augmentation Phrases

Table 6 lists our manually curated list of phrases inserted during data augmentation. All phrases were derived via a small manual error analysis on the `Base` model.

We broadly divide our phrases into three categories: introductory statements, hedging, and sentiment statements. For each instance in ANLI, one random phrase from the list is prepended to the hypothesis. We use all three rounds of ANLI annotations. This results in 162,865 augmented instances which, together with the original ANLI instances, leads to a total of 325,730 training instances.

Parameter	Val.
Warmup Ratio	0.06
Weight Decay	0.01
Effective Batch Size	64

Table 7: Hyperparameters

A.2 Hyperparameters

Table 7 lists the hyperparameter settings for our model. We use the same optimizer hyperparameters as Laurer et al. (2022) except for an increased batch size and the learning rate. For the latter we tested three learning rates ($5e-6$, $5e-2$, $5e-1$) and select the one that provided the best loss on the augmented ANLI validation set. We initially ran models for 10,000 steps with a checkpoint every 1,000 steps and selected the checkpoint with the lowest loss on the augmented ANLI validation set. Later we reduced the number of training steps to 2,000 since we found we would usually select an early checkpoint as validation loss increased later in training, likely related to overfitting on the augmented data.

A.3 Training

We use the DeBERTa implementation in the huggingface transformers library (Wolf et al., 2020) and trained our model on a single node using two RX6800 GPUs, with one training run taking about three hours. Later experiments with fewer steps cut that time by 80%.

B Dataset Bias in BEGIN

BEGIN is the only dialogue corpus on which first person pronoun occurrence shows a strong (negative) correlation with faithfulness (see Table 3). Since there is nothing in the annotation guidelines that would explain this correlation, we instead hypothesize that this is the consequence of a model induced bias in the data. Specifically, we hypothesize that one of the two models in BEGIN is (1) *more* likely to generate personal statements and (2) *less* likely to generate faithful responses.

To avoid confusion in the remainder of this section, we highlight that there are two variants of BEGIN:

BEGIN-v1 is the variant used in TRUE. It contains labeled generations by a fine-tuned GPT-

2 base (Radford et al., 2019) and a fine-tuned T5 base model (Raffel et al., 2020) on the Wizard of Wikipedia dataset (Dinan et al., 2019).[6]

BEGIN-v2 is a more recent variant of BEGIN that is not part of TRUE. In addition to *new* instances generated by T5 and GPT-2 it contains outputs from two additional models. It also has a revised annotation procedure. When we refer to BEGIN-v2, we exclusively mean the Wizard of Wikipedia subset.

Unfortunately, BEGIN-v1 does not allow us to retrieve which model generated which instance. This makes it impossible to directly investigate for model bias. However, BEGIN-v2 includes outputs by the same two models, fine-tuned on the same data. Since we only need corpus level statistics to verify our assumptions, we conduct our analysis on the GPT-2 and T5 instances in BEGIN-v2.

To verify (1), we compute the correlation between a binary variable indicating which model generated each instance (T5: 0, GPT-2: 1) and first-person pronoun occurrence. We find a positive correlation (Kendall's τ wrt. to *I*-pronoun occurrence: 0.18, $p < 0.001$), indicating that GPT-2 generates outputs including more first-person pronouns.

To investigate whether GPT-2 is also more likely to be unfaithful, i.e. to verify (2), we compute the correlation between the binary model indicator variable and a faithfulness variable that is 1 when the output is labelled as *Fully attributable* and 0 otherwise. We find a negative correlation (Kendall's τ wrt. to Faithfulness: -0.25, $p < 0.001$), supporting our hypothesis that GPT-2 is also overall less faithful. To ensure that this is not an effect of additional personal statements leading to more unfaithful generations, we conduct the same analysis only on instances where we identify no first-person pronouns. We find a similarly strong negative correlation of -0.29 ($p < 0.001$).

Our analysis shows that GPT-2 produces both overall less faithful outputs and more first-person pronouns than T5. Since BEGIN-v1 contains only outputs from T5 and GPT-2 this suggests that the root cause for the negative correlation between faithfulness label and first-person pronoun occurrence in BEGIN-v1 is model bias confounding faithfulness and first-person pronoun occurrence.

[6]The relevant data can be found at https://raw.githubusercontent.com/google/BEGIN-dataset/5fa0cb0dde0e653d2016724a52a5ca27fe8b6a3f/dev_05_24_21.tsv

Corpus	Faith.	Non. Faith	Total
Frank	223 (33.2%)	448 (66.8%)	671
MNBM	255 (10.2%)	2245 (89.8%)	2500
SummEval	1306 (81.6%)	294 (18.4%)	1600
QAGS-X	116 (48.5%)	123 (51.5%)	239
QAGS-C	113 (48.1%)	122 (51.9%)	235
BEGIN	282 (33.7%)	554 (66.3%)	836
DialFact	3341 (38.5%)	5348 (61.5%)	8689
Q2	628 (57.7%)	460 (42.3%)	1088
PAWS	3539 (44.2%)	4461 (55.8%)	8000

Table 8: Dataset statistics for all constituent corpora in TRUE

B.1 Dataset Bias in BEGIN-v2

We conduct a preliminary study to investigate whether similar biases also exist in BEGIN-v2.

We observe that while BEGIN-v2 uses data from four dialogue systems, a majority of faithful generations is produced by a single system called CTRL-DIALOG (Rashkin et al., 2021). CTRL-DIALOG is specifically trained to generate less subjective text, which we hypothesize might result in fewer first person pronouns. Since CTRL-DIALOG also produces more faithful texts, this would lead to a negative correlation between faithfulness and first person pronouns, similar to what we observe on BEGIN-v1.

We verify this assumption by computing the correlation of a binary variable indicating an instance has been generated by CTRL-DIALOG with a) the faithfulness labels on BEGIN-v2 and b) first-person pronoun occurrence. We find that an instance being generated by CTRL-DIALOG is positively correlated with it having a *faithful* label (Kendall τ w.r.t. faithfulness: 0.48, p$<$ 0.001) while being negatively correlated with the number of pronouns (Kendall τ w.r.t. *I*-pronoun occurrence: -0.34, p$<$ 0.001). This suggests future evaluations on the BEGIN-v2 might run into similar bias issues.

C Dataset Statistics

We report the number of instances, as well as the class distribution of TRUE in Table 8.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
8

☑ A2. Did you discuss any potential risks of your work?
9

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

1,3

☑ B1. Did you cite the creators of artifacts you used?
1,3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
1,9

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
9

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Most data is machine generated and thus unlikely to reveal personal information. All data is also already publicly available and has been introduced in peer-reviewed publications, providing an additional safeguard.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We discuss the limitation to English in Section 9.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix C

C ☑ Did you run computational experiments?

3,4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
5.2,Appendix A

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.