Less is More for Improving Automatic Evaluation of Factual Consistency

Tong Wang, Ninad Kulkarni, Yanjun Qi
AWS Bedrock Science
{tonwng, ninadkul, yanjunqi}@amazon.com

Abstract

Assessing the factual consistency of automatically generated texts in relation to source context is crucial for developing reliable natural language generation applications. Recent literature proposes AlignScore which uses a unified alignment model to evaluate factual consistency and substantially outperforms previous methods across many benchmark tasks. In this paper, we take a closer look of datasets used in AlignScore and uncover an unexpected finding: utilizing a smaller number of data points can actually improve performance. We process the original AlignScore training dataset to remove noise, augment with robustness-enhanced samples, and utilize a subset comprising 10% of the data to train an improved factual consistency evaluation model, we call LIM-RA (`Less Is More for Robust AlignScore`). LIM-RA demonstrates superior performance, consistently outperforming AlignScore and other strong baselines like ChatGPT across four benchmarks (two utilizing traditional natural language generation datasets and two focused on large language model outputs). Our experiments show that LIM-RA achieves the highest score on 24 of the 33 test datasets, while staying competitive on the rest, establishing the new state-of-the-art benchmarks.

1 Introduction

The emergence of large language models (LLMs) and an increasing interest in utilizing machine-generated texts from like summarization, paraphrasing, and question-answering (QA) has created a need to automatically evaluate the degree to which generated natural language texts accurately reflect the factual information contained in source context. Early work used Natural Language Inference (NLI) (Laban et al., 2022) and QA (Fabbri et al., 2021) to handle automatic factual consistency evaluation. However, these methods exhibit limited generalizability and struggle with handling long contexts. Recently, Zha et al. (2023) propose AlignScore, a unified model based on RoBERTa and is trained on a wide range of datasets to calculate the alignment between context and generated text. AlignScore achieves state-of-the-art results across several factual consistency benchmarks.

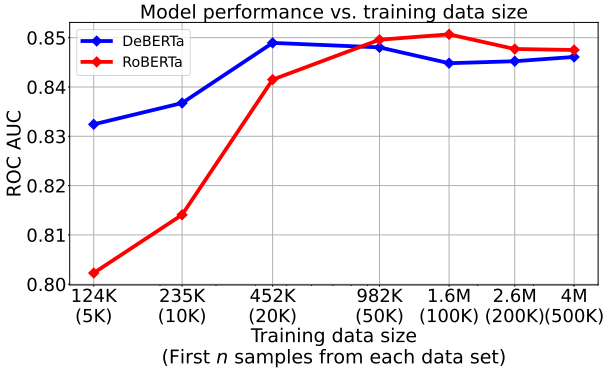

Figure 1: Ablation study on using the first n samples from each sub-train dataset for training and overall model performance. We see that the optimum benchmark performance is 452K and 1.6M samples for DeBERTa and RoBERTa respectively. For comparison AlignScore uses 4.7M or the first 500K. Performance broken down by benchmark can be found in A.1

Despite its strengths, the AlignScore study has several limitations. First, the training data leveraged for developing AlignScore were derived in a heuristic manner from many existing NLP tasks and datasets, adding noise and poor quality in some samples. We, therefore, ask the question: Are all data points from AlignScore training needed? Our ablation studies shown in Figure 1 indicate that the answer is "No". Additionally, AlignScore displays fragility regarding robustness, as it fails to identify some clear perturbations involving entities like names, numbers, etc. As Table 1 illustrates, even simple modifications can produce false positives and false negatives when using AlignScore.

In this paper, we propose **LIM-RA** (**L**ess **I**s **M**ore - **R**obust **A**lignScore), an improved version

Context	Claim	AlignScore	LIM-RA
[...] Napoleon married the Archduchess Marie Louise, who was 18 years old [...]	Archduchess Marie Louise was 18 years old when she married Napoleon .	0.9907	0.9542
	Archduchess Mari Louze was 18 years old when she married Napoleon .	0.9650 (false positive)	0.4381
The Blue Ridge Mountains [...] attain elevations of about 2,000 ft	The typical elevations of the Blue Ridge Mountains are 2,000 ft.	0.9812	0.9434
	The typical elevations of the Blue Ridge Mountains are 2000 ft.	0.0214 (false negative)	0.8621

Table 1: Examples of robustness issues in AlignScore predictions. In the first example we perturb the correct name "Marie Louise" to the incorrect name "Mari Louze"; however, the factual consistency score is still high, resulting in a false positive. Similarly in the second example we perturb "2,000" to "2000", resulting in a false negative.

of AlignScore trained on DeBERTa (He et al., 2021). Our model is the result of multiple ablation steps on improving training data quality, analyzing training size as well as constructing synthetic data to improve robustness (Figure 2 shows overall workflow). We demonstrate that with about 10% of the cleaned training data, we are able to obtain a better model than AlignScore. Our experiments show that LIM-RA consistently outperforms strong baselines including AlignScore and GPT-3.5-Turbo, achieving the new state-of-the-art on four factual consistency benchmarks covering a wide range of 33 datasets. It is worth noting that our experiments include a newly defined benchmark, Large Language Model Response (LLMR), designed for evaluating LLM outputs' factual consistency. LIM-RA performs the best on LLMR.

2 Method

2.1 AlignScore Model and Training Data

Automatic evaluation of factual consistency is challenging. Recently proposed AlignScore measures the alignment of information between machine-generated natural language texts and given source material to evaluate the factual accuracy (Zha et al., 2023). AlignScore is built on top of a unified alignment function via RoBERTa (Liu et al., 2019) and trained on datasets derived from 7 NLP tasks: NLI, QA, Fact Verification, Paraphrase, Semantic Textuality Similarity, Information Retrieval, and Summarization. Each sample in a task is converted into a text pair (context, claim) and a label. The label has 3 options based on the task and dataset: binary (aligned, not-aligned), 3-way (aligned, contradict, neutral), regression (score between 0 to 1). For example in SNLI dataset, the context is the premise, the claim is the hypothesis, label is the 3-way label. Certain prepossessing steps are required to unify the format in multiple datasets.

To calculate the factual consistency score of long text, AlignScore first splits the context into roughly 350-token chunks and the claim into sentences. Then the trained alignment function (RoBERTa based) evaluates each sentence in the claim against each context chunk. For example, in the 3-way classification head, the probability of the "aligned" class is used as the alignment score. The highest alignment score for each claim sentence is selected and then averaged to obtain the overall factual consistency score. By using the chunking strategy, AlignScore can be applied to text of any length, as shown by Figure 3.

2.2 Training Data Cleaning

For training, AlignScore uses more than 30 datasets and selects 500K samples from each dataset to build its trainng data, including a total of 4.7M training samples. Training the AlignScore alignment model requires 5 days on 8 V100 GPUs.

However, we find that not all the training datasets have good quality. The upper half of Figure 2 shows a cohort of data cleaning steps we use to improve the training data quality. First, based on our ablation studies, we remove four datasets that do not result in performance gains, such as ms_marco and wikihow. Additionally to prevent the model from truncating sentences that support the claim, we only keep samples in which the context has fewer than 512 tokens.

When using QA datasets to create alignment training samples, since the QA passage is the context, a preprocessing step is needed. AlignScore uses a pre-trained sequence-to-sequence model to convert question-answer into a declarative sentence as the input claim. We, however, observed a performance decrease in our experiments when using this preprocessing. We find the decrease was because the generated declarative sentence has poor data quality. Thus, we concatenate question and answer as the claim text. [1]

Additionally, many QA datasets only have

[1]We also tried to use Mistral-7B (Jiang et al., 2023) few-shot to generate better-quality declarative sentences but still did not produce performance gains.

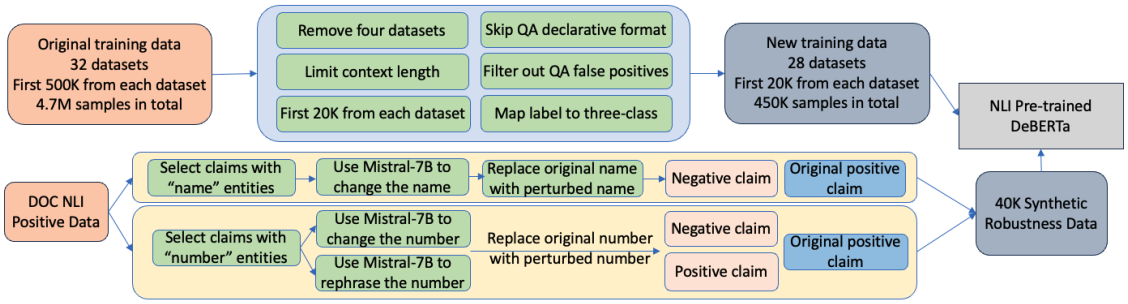

Figure 2: Overall workflow of our method is depicted in the diagram. The top workflow describes how we clean the training data, the bottom workflow illustrates the process of creating synthetic robustness data. Then we train a pre-trained DeBERTa model on those data to obtain LIM-RA.

ground truth answers (positive samples) but no wrong answers (negative samples). To address this, AlignScore generates fake wrong answers using the T5 model, and answers the question based on the original passage with the ground truth answer tokens masked. However, this leads to false negatives because many generated fake answers are similar to or exactly match their corresponding ground truth answers. To mitigate the issue, we use Sentence-BERT (Reimers and Gurevych, 2019) to encode both the fake and ground truth answers, and then filter out the fake answers that are similar to the true answers by using rules and a threshold of 0.85. This data cleaning procedure is illustrated in the top half of figure 2.

After cleaning the data, we use 20K samples from each dataset for a total of 452K training samples (about 10% of training data used for AlignScore) which results in a better model (results in Section 3.3).

2.3 Synthetic Robustness Data

We also notice AlignScore fails on name or number perturbations as illustrated in Table 1. To mitigate the issue, we augment the training dataset by creating a synthetic dataset designed to enhance the model's robustness, with emphasis on name and number variation based text generation as illustrated in the bottom half of figure 2.

We create two synthetic datasets: Robust-Name and Robust-Number datasets using DocNLI (Yin et al., 2021). DocNLI includes multiple-sentence contexts and single-sentence claims discussing facts in the context. To create the Robust-Name data, we use spaCy NER (Honnibal and Montani, 2017) to identify the "PERSON" and "ORG" entities in samples labeled as "entailment" and use Mistral-7B to perturb the entities (prompt details in Appendix A.3). The original entity is replaced with the perturbed entity to construct the synthetic negative samples. Using Mistral instead of randomly perturbing a character in the entity ensures the new name is similar to a real person or org name. The two-step generation generates a better rewritten claim than directly instructing the LLM to rewrite the claim.

Similarly, we construct the Robust-Number data by perturbing claims with number-related labels such as "TIME", "QUANTITY", "DATE". We use Mistral to rephrase ("100" to "one hundred") and change numbers ("100" to "101"). The perturbed entities replace the original to create positive and negative data.

2.4 LIM-RA Model

We experiment with different pretrained models as base including RoBERTa (large), DeBERTa (large), DistilBERT (base). DeBERTa achieves the best overall performance while DistilBERT has poor performance due to its small model capacity. Also, we unify all data labels to the three class setup (details later in this section), and use the 3-way classification head to predict *aligned* (factual consistent), *neutral* (no-evidence), and *contradiction*. At inference time, we follow AlignScore to split context into chunks and claim into sentences, and average the sentence alignment scores to compute the overall factual consistency score. We denote LIM-RA and LIM-A as the DeBERTa model trained with cleaned data and with and without synthetic robustness in training, respectively.

Under the Hood: We train a pre-trained NLI DeBERTa model[2] (Laurer et al., 2024) for 3 epochs using AdamW optimizer with learning rate as 1e-5.

[2] *https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli*

We use the first 20k samples from each of the 28 train datasets described in AlignScore, plus the 2 new synthetic robustness datasets, resulting in a total of 490k samples in our final training. Hyperparameter details can be found in Table 10. We follow AlignScore and use the factual consistency class probability as the alignment score.

Unifying Labels: We convert binary and regression labels to 3-class labels. For datasets with binary labels, we map the negative label "not-aligned" to either "contradiction" or "no-evidence" depend on the dataset. In most of the cases, we map the negative label to "contradiction", such as in *doc_nli* and *paws*. But in *qqp*, we map the negative label to "no-evidence". For regression labels in *stsb* dataset, we bin the score as three classes: faithful ($>= 0.45$), no-evidence ($>= 0.3, < 0.45$), contradiction (< 0.3).

2.5 Connecting to Related Works

Previous studies include multiple other methods for assessing factual consistency. (1) QA-based factual consistency, including QuestEval (Scialom et al., 2021) and QAFactEval (Fabbri et al., 2021), checks if the source answer is different from the target answer given a question. (2) With the recent advances in LLMs, a new line of research is to evaluate factual consistency directly with an LLM (Liu et al.; Fu et al., 2023a; Jia et al., 2023). (Chen et al., 2023) investigate a variety of prompting methods including vanilla prompting, chain-of-thought prompting, and a sentence-by-sentence prompting and (Luo et al., 2023) explore ChatGPT's ability to evaluate factual inconsistency under a zero-shot setting while (Fu et al., 2023b) uses LLMs in a QA setting for direct factual consistency scoring. (3) A third related line of methods uses the Natural Language Inference (NLI) based formulation. For instance (Laban et al., 2022) proposed SummaCConv, that segments documents into sentences and aggregates NLI scores between pairs of sentences.

Factual consistency benchmark datasets typically contain (context, claim, label) triplets where the label indicates if the claim is consistent with the context and is difficult to obtain as high-quality annotation is challenging due to low inter-annotator agreement (Falke et al., 2019; Laban et al., 2022). (Laban et al., 2022) introduce the SummaC (Summary Consistency) benchmark which consists of 6 large inconsistency detection datasets standardized as a binary classification task given document and summary. (Laban et al., 2023) introduce SummEdits, a summarization consistency dataset where an LLM introduces inconsistencies in an otherwise consistent summary and show that the benchmark is challenging for most current LLMs. (Honovich et al., 2022) present TRUE, which consolidates 11 existing datasets covering summarization, knowledge-grounded dialogue, paraphrasing and fact verification annotated for consistency.

Model	CG	XF	FC	SE	FRK	AVG
NER	54.4	69.0	50.8	59.3	68.4	60.4
Questeval	59.7	65.6	73.3	76.9	86.3	72.4
QAFactEval	82.5	65.1	89.2	88.5	89.6	82.9
SummaC	65.6	70.3	92.2	86.0	88.4	80.5
AlignScore	76.9	**78.1**	89.1	82.3	88.0	82.9
LIM-A	84.2	73.9	**93.7**	**92.4**	90.3	86.0
LIM-RA	**84.9**	75.7	93.2	92.2	**92.0**	**87.6**

Table 2: SummaC benchmark AUC-ROC results. LIM-RA and LIM-A outperform all current baselines in 4 of the 5 datasets with LIM-RA performing the best overall.

3 Experiments

We conduct a comprehensive experimental study to evaluate LIM-RA on multiple factual consistency benchmarks and demonstrate LIM-RA consistently outperforms strong baselines and establishes new state-of-the-art results. Our experiments also include ablation studies (Table 7) and robustness analysis (Table 9) of LIM-RA. We list the hyperparameters we used for LIM-RA in Table 10. Each of our experiments covers 20 different random seeds.

3.1 Four Benchmarks: 33 Datasets

We evaluate the factual consistency performance using AUC-ROC on 33 datasets from 4 benchmarks: SummaC, SummEdits, TRUE, and LLMR. Each data sample in the benchmarks is a pair of target text (claim) and a grounding source text (context), with a binary annotation of whether the target text is factually consistent w.r.t its source. The benchmark dataset details can be found in Appendix A.2.

SummaC 5 summary consistency datasets: GoGenSumm (CG), XsumFaith (XF), FactCC (FC), SummEval (SE), Frank (FRK). We remove Polytope dataset since it contains negative samples that do not imply factual consistency errors.

TRUE 11 datasets covering summarization, knowledge-grounded dialogue, paraphrasing and fact verification annotated for factual consistency: Frank (FRK), SummEval (SE), MNBM, QAGS-

Model	ECT	QM	SCall	SS	SCI	SEmail	NEWS	BILL	PD	SP	AVG
NER	59.7	55.2	56.3	57.6	53.3	66.8	60.9	49.1	57.8	51.5	56.8
Questeval	64.8	54.0	63.1	54.4	51.9	55.9	64.9	59.8	54.1	53.4	57.6
QAFactEval	75.8	65.5	74.6	71.3	69.7	69.8	81.4	56.9	64.0	65.5	69.4
SummaC	66.7	55.8	61.1	54.3	61.0	58.9	61.1	54.5	61.0	61.1	59.6
AlignScore	91.5	83.8	89.1	**85.5**	82.1	81.6	80.6	61.6	78.0	72.3	80.6
LIM-A	**93.6**	86.9	90.7	83.1	**87.7**	**82.5**	**82.1**	69.3	81.0	86.9	84.4
LIM-RA	92.8	**88.2**	**91.2**	84.2	86.0	81.1	81.3	**72.1**	**82.5**	**86.9**	**84.6**

Table 3: SummEdits benchmark AUC-ROC results. LIM-RA and LIM-A are the best performing models in 9 of the 10 datasets and LIM-RA is the best performing model overall.

Model	ATS	BBA-4	BBA-16	BBS-4	BBS-16	PHD	HE	AVG
AlignScore	62.7	62.4	59.4	71.8	75.2	74.6	73.1	68.5
LIM-A	65.9	69.2	60.4	79.5	78.4	**78.0**	72.2	71.9
LIM-RA	**66.3**	**71.4**	**60.7**	**82.5**	**79.6**	77.0	**74.9**	**73.2**

Table 4: LLMR benchmark AUC-ROC results. We compare against only AlignScore as it is the best performing baseline as seen in tables 2 and 3. LIM-RA and LIM-A are the best performing models in all 7 datasets. LIM-RA is the best performing model overall.

Model	BEGIN	DF	FVR	FRK	MNBM	PAWS	Q^2	QC	QX	SE	VITC	AVG*	AVG
NER	50.6	62.7	62.4	65.5	68.4	51.7	59.1	48.4	63.6	56.6	57.8	59.3	58.8
Questeval	**83.9**	77.2	72.5	84.0	64.8	69.0	72.2	64.5	55.2	69.7	66.6	71.4	70.9
QAFactEval	81.0	81.8	86.0	88.5	67.3	86.1	75.8	83.9	76.1	80.9	73.6	79.4	80.1
SummaC	81.6	81.2	92.0	89.0	67.2	88.2	77.5	77.7	76.0	79.1	97.5	78.7	82.5
AlignScore	81.4	85.0	94.9	88.7	**78.2**	98.3	79.1	**89.6**	**83.1**	71.4	**98.4**	82.0	86.2
LIM-A	79.0	**85.2**	**95.5**	90.0	74.7	98.4	**83.5**	85.0	82.4	83.5	97.1	83.0	86.8
LIM-RA	80.8	83.8	95.2	**91.3**	75.8	**98.4**	82.7	84.5	82.7	**84.8**	96.8	**83.3**	**87.0**

Table 5: TRUE benchmark AUC-ROC results. LIM-RA and LIM-A are the best performing models in 6 of the 11 datasets. LIM-RA is the best performing model overall. We report AVG* in the second last column by excluding PAWS, FVR, and VITC to show out-of-domain performance.

CNNDM (QC), QAGS-Xsum (QX), BEGIN, Q^2, DialFact (DF), Fever (FVR), VitaminC (VITC), PAWS.

SummEdits 10 datasets evaluating factual consistency in summarization covering multiple domains. Inconsistent summaries are generated by GPT-3.5-Turbo: News, Podcast (PD), Billsum (BILL), Samsum (SS), Shakespeare (SP), SciTLDR (SCI), QMSum (QM), ECTSum (ECT), Sales Email (SEmail), Sales Call (SCall).

LLMR (large language model response) is a new benchmark consisting of 7 datasets we introduce in this paper. Similar to SummEdits, the datasets are designed to evaluate the factual consistency of LLM output and inconsistencies are generated in an automated fashion with human verification: HaluEval (HE) (Li et al., 2023) consists of CNN/DailyMail articles with correct and hallucinated summaries generated by ChatGPT in a zero-shot manner. BAMBOO abs-hallu (BBA) and sen-hallu (BBS) subsets (Dong et al., 2023) consist of NLP academic papers (max 4K and 16K token variants for a total of 4 datasets) with supported and hallucinated hypotheses generated by ChatGPT similar to HE. Passage-level Hallucination Detection (PHD) (Yang et al., 2023) consists of Wikipedia articles of an entity with correct and hallucinated biographies of that entity generated by ChatGPT. AttrScore (ATS) (Yue et al., 2023) consists of QA datasets and New Bing search queries in the format $(question, answer, context, label)$ where $label$ indicates if the $answer$ is supported by $context$. Hallucinations are generated by both swapping the answer with an incorrect answer and by swapping the the context with another article. For our experiments we consider context as $document$ and answer as $claim$.

3.2 Baselines Methods

NER (Laban et al., 2022), uses spaCy NER to match entities between claim and context.

Questeval, QA-based model, evaluates both factual consistency and relevance of the generated text by checking if the answer from source is different from the answer from target given a question.

QAFactEval, QA-based model, evaluates factual consistency by performing answer selection, question generation, question answering, and answer

Model	SummaC	TRUE	SummEdits	LLMR	AVG
AlignScore	82.9	86.2	80.6	68.5	79.6
LIM-A	86.0 (+3.7%)	86.8 (+0.7%)	84.4 (+4.7%)	71.9 (+5.0%)	82.3 (+3.4%)
LIM-RA	**87.6 (+5.7%)**	**87.0 (+0.9%)**	**84.6 (+5.0%)**	**73.2 (+6.9%)**	**83.1 (+4.4%)**

Table 6: Average AUC results and relative improvements over AlignScore on four benchmarks. The last column is the overall average of SummaC, TRUE, SummEdits, and LLMR scores.

Model	Setting	Overall
AlignScore	4.7M	83.2
RoBERTa	pre-train	71.6
DeBERTa	pre-train	82.1
RoBERTa	10% + cleaning	84.1
DeBERTa	10% + cleaning	83.6
RoBERTa	10% + cleaning + pre-train	83.8
LIM-A	10% + cleaning + pre-train	86.0
LIM-RA	+syn robust data	86.4

Table 7: Ablation Study

overlap evaluation.

SummaC, NLI-based model (SummaCConv), segments documents into sentence units and aggregates scores between pairs of sentences.

AlignScore, current state-of-the-art, an alignment function trained on a wide range of datasets.

0-shot/10-shot GPT-3.5-Turbo, instruct the LLM to evaluate whether the claim is consistent, lacks evidence, or contains contradictions.

10-shot Mistral-7B, one of the best performing open-source LLMs. We use the same prompts as 10-shot GPT-3.5-Turbo.

3.3 Experimental Results

3.3.1 Results on Traditional Benchmarks: SummaC and TRUE

We evaluate factual consistency models on the SummaC benchmark in Table 2. LIM-RA achieves the best overall score and has a 5.7% relative improvement over AlignScore and QAFactEval. Our model has the top result in 4 of the 5 datasets. Our results for AlignScore are lower than the results reported in the original work (Zha et al., 2023) because we did not include the rule-based inference-time processing (such as removing special tokens or capitalizing the first letter) for a fair comparison between all models.

From the results on the TRUE benchmark in Table 5, we see that LIM-RA has the best overall AUC-ROC score with a 0.9% improvement over AlignScore and has the best score in 5 of 11 datasets. As suggested in (Zha et al., 2023), we report AVG* by removing PAWS, FVR, and VITC to show out-of-domain performance; LIM-RA remains the best performing model.

3.3.2 Results on LLM output: SummEdits and LLMR

We evaluate factual consistency on LLM responses using the SummEdits and LLMR benchmarks in Table 3 and Table 4 respectively. On the SummEdits benchmark, both LIM-A and LIM-RA consistently outperform other baselines. LIM-RA has the best overall performance and has a 5.0% relative improvement over the best baseline AlignScore. Our model achieves the best score in 8 of the 10 datasets and performs significantly better on OOD domain datasets such as Shakespeare (SP), BillSum (BILL), SciTLDR (SCI) compared to the baseline. On the LLMR benchmark, we only report AlignScore as Tables 2, 3, 5 show that AlignScore is the strongest baseline. LIM-RA achieves the best overall result and obtains a relative improvement of 6.9% over AlignScore, and has the best score on 6 of the 7 datasets.

We report the overall average score on the four benchmarks in Table 6. In summary, LIM-RA exhibits a 4.4% relative improvement over the baseline model AlignScore.

3.3.3 Comparing with LLM Baselines

We compare the trained metric models with two LLMs: Mistral-7B and GPT-3.5-Turbo (ChatGPT) using the same 0-shot and 10-shot prompt (described in Appendix A.4). Since LLMs do not provide factual consistency scores, we report balanced accuracy in Table 8 and only report SummaC and SummEdits due to time constraints. LIM-RA continues to perform the best on the two benchmarks while GPT-3.5-Turbo outperforms Mistral by a large margin on SummaC. Additionally, 0-shot ChatGPT outperforms 10-shot ChatGPT on SummEdits possibly because the 10-shot demonstrations are out-of-domain. We compare average inference time of each model on a sample of data from SummaC and find AlignScore demonstrates fast inference speed of 0.18s on a single NVDIA-A10G GPU followed by LIM-RA with 0.29s. The slower speed is because DeBERTa is slower than RoBERTa even though they have a similar number of parameters. 0-shot ChatGPT and Mistral-7B on

Model	SC	SE	Time	GPUs
Mistral-7B 10-shot	62.0	64.0	0.51s	4
GPT-3.5 0-shot	73.5	71.6	0.52s	API
GPT-3.5 10-shot	76.7	69.8	8.4s	API
AlignScore	74.0	71.9	0.18s	1
LIM-A	77.8	76.5	0.29s	1
LIM-RA	**78.5**	**76.7**	0.29s	1

Table 8: Evaluation using LLMs Balanced Accuracy results and Average Inference Time on SummaC (SC) and SummEdits (SE).

	Robust-Name	Robust-Number
Train (Test)	19,508 (3,492)	20,628 (5,076)
AlignScore	64.3	86.4
LIM-A	64.0	88.0
LIM-RA	**84.8**	**91.8**

Table 9: Synthetic robustness data size and AUC-ROC performance across models when facing perturbed data.

4 GPUs using vLLM (Kwon et al., 2023) achieves comparable speed of 0.52s and 0.51s respectively while OpenAI GPT-3.5 10-shot is the slowest, primarily due the to the rate limit of a Tier-1 account[3].

3.4 Results on Synthetic Robustness Data

In Table 9 we evaluate the models on the synthetic robustness test dataset created in section 2.3. We see LIM-A without synthetic data augmentation performs on par with AlignScore while LIM-RA performs the best and is more robust to name and number perturbations.

3.5 Ablation Analysis

We perform ablation studies to answer the following questions: (1) What is the impact of different training data sizes? (2) What is the performance of using a pre-trained model as the alignment? (3) What is the impact of the cleaned data? and (4) What is the impact of fine-tuning RoBERTa or DeBERTa as the alignment function?

To answer (1) we sweep the size from 123K (5K per dataset) to 4M (500K per dataset). From Figure 1, we see that the benchmark performance peaks at 452K and 1.6M samples for DeBERTa and RoBERTa respectively and reduces if we include more data. For (2)-(4), we report the average AUC-ROC score of SummaC, SummEdits, TRUE in Table 7. To answer (2), we experiment with different off-the-shelf pre-trained NLI models. The best pre-trained DeBERTa model (82.1%) outperforms the best pre-trained RoBERTa[4] (71.6%) (Nie et al., 2019). To answer (3), we perform data cleaning and use 10% (452K samples) of the training data and find both RoBERTa (84.1%) and DeBERTa (83.6%) outperform AlignScore (83.2%). To answer (4), we fine-tune the pre-trained models using the cleaned data. DeBERTa (LIM-A) performance improves with fine-tuning while RoBERTa performance decreases, possibly because the pre-trained DeBERTa outperforms the pre-trained RoBERTa model. Finally, adding the synthetic robustness data can further boost the performance.

[3]The Tier-1 rate-limit for GPT-3.5-Turbo is 60K tokens per minute, 3.5K requests per minute, and 10K requests per day. https://platform.openai.com/docs/guides/rate-limits/usage-tiers?context=tier-one

4 Conclusions

We propose LIM-RA, a DeBERTa based model to automatically evaluate factual consistency trained from a cleaner and smaller training set than used for AlignScore. Experimental results show LIM-RA consistently outperforms the current state-of-the-art AlignScore and other strong baselines on 4 benchmarks. In addition, the model is robust to name and number variations and is better suited for LLM outputs' factual consistency evaluation.

References

Shiqi Chen, Siyang Gao, and Junxian He. 2023. Evaluating factual consistency of summaries with large language models. *arXiv preprint arXiv:2305.14069*.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models. *arXiv preprint arXiv:2309.13345*.

Alexander R Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021. Qafacteval: Improved qa-based factual consistency evaluation for summarization. *arXiv preprint arXiv:2112.08542*.

Tobias Falke, Leonardo FR 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.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023a. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Xue-Yong Fu, Md Tahmid Rahman Laskar, Cheng Chen, and Shashi Bhushan TN. 2023b. Are large

[4]*https://huggingface.co/ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli*

language models reliable judges? a study on the factuality evaluation capabilities of llms. *arXiv preprint arXiv:2311.00681*.

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

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

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. *arXiv preprint arXiv:2204.04991*.

Qi Jia, Siyu Ren, Yizhu Liu, and Kenny Q Zhu. 2023. Zero-shot faithfulness evaluation for text summarization with foundation language model. *arXiv preprint arXiv:2310.11648*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Philippe Laban, Wojciech Kryściński, Divyansh Agarwal, Alexander R Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. 2023. Llms as factual reasoners: Insights from existing benchmarks and beyond. *arXiv preprint arXiv:2305.14540*.

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, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment, may 2023. *arXiv preprint arXiv:2303.16634*.

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

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. 2021. Questeval: Summarization asks for fact-based evaluation. *arXiv preprint arXiv:2103.12693*.

Shiping Yang, Renliang Sun, and Xiaojun Wan. 2023. A new benchmark and reverse validation method for passage-level hallucination detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3898–3908, Singapore. Association for Computational Linguistics.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. Docnli: A large-scale dataset for document-level natural language inference. *arXiv preprint arXiv:2106.09449*.

Xiang Yue, Boshi Wang, Kai Zhang, Ziru Chen, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. *arXiv preprint arXiv:2305.06311*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*.

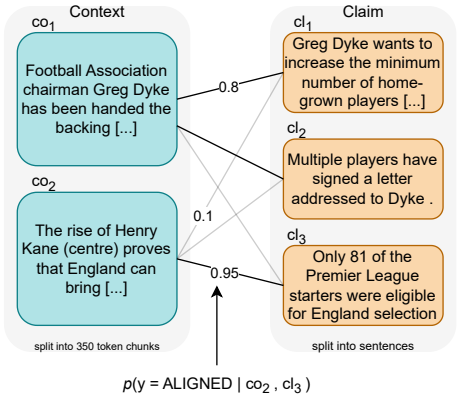

Figure 3: Visual description of AlignScore. The context and claim are split into 350 token and sentence chunks respectively. Then an alignment function evaluates each *(context chunk, claim sentence)*. The factual consistency score is calculated by first selecting the highest alignment score for each *claim* and then averaging these scores across all *claims*.

A Appendix

A.1 Training data size ablation for each benchmark dataset

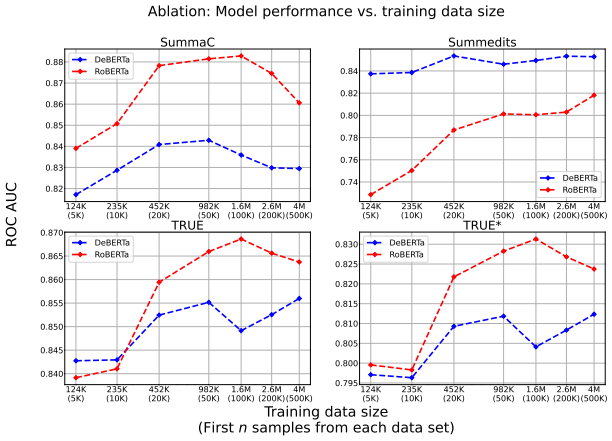

Figure 4: Ablation study on using the first n samples for training and model performance on each benchmark data set.

Parameter	Value
samples_per_dataset	20000
max_context_length	512
lr	1e-5
seed	2027
train_batch	8
accumulate_grad_batch	1
epoch	3
warmup_ratio	0.06
weight_decay	0.01
adam_epsilon	1e-6

Table 10: LIM-RA Hyperparameters

A.2 Benchmark Details

Tables 11, 12, 13, 14 describe the total number and number of factually consistent samples in each benchmark and dataset.

Dataset	# Samples	# Factually Consistent
CG	400	312
XF	1250	130
FC	503	441
SE	850	770
FRK	1575	529
Total	4578	2182

Table 11: SummaC Benchmark

Dataset	# Samples	# Factually Consistent
ECT	668	242
QM	431	183
SCall	520	173
SS	664	242
SCI	466	145
SEmail	613	179
NEWS	819	321
BILL	853	361
PD	500	163
SP	814	378
Total	6348	2387

Table 12: Summedits Benchmark

Dataset	# Samples	# Factually Consistent
BEGIN	836	282
DF	8689	3341
FVR	18209	6393
FRK	671	223
MNBM	2500	255
PAWS	8000	3539
Q^2	1088	628
QC	235	113
QX	239	116
SE	1600	1306
VITC	63054	31484
Total	105121	47680

Table 13: TRUE Benchmark

Dataset	# Samples	# Factually Consistent
ATS	4241	1414
BBA-4	200	100
BBA-16	200	100
BBS-4	200	100
BBS-16	200	100
PHD	299	222
HE	20000	10000
Total	25340	12036

Table 14: LLMR Benchmark

A.3 Few-shot Prompt to Generate Synthetic Robustness Data

A.3.1 Prompt to perturb names

```
Given a name, modify one or two
letters to change it to a different
name.

Original Text: Abraham Lincoln
Changed Text: Abrahem Lincoln

Original Text: cricket
Changed Text: cracket

Original Text: Wireshark
Changed Text: Wileshark

Original Text: Robert Urquhart.
Changed Text: Robert Uruhart.

Original Text: Dee Smith
Changed Text: Dee Smyth

Original Text: Emma Wastson
Changed Text:
```

A.3.2 Prompt to perturb numbers

```
Change the meaning of the text.

Original Text: 37
Changed Text: 27

Original Text: more than 10
years ago
Changed Text: more than 11
years ago

Original Text: more than 10
years ago
Changed Text: within 10 years

Original Text: second
Changed Text: third

Original Text: 22 June 1990
Changed Text: 22 July 1990

Original Text: at least one
Changed Text: at most one

Original Text: 2 years
Changed Text:
```

A.3.3 Prompt to rephrase numbers

```
Rephrase the numbers in the text.

Original Text: 154
Rephrase Text: one hundred
fifty-four

Original Text: more than 10
years ago
Rephrase Text: more than ten
years ago

Original Text: second
Rephrase Text: 2nd

Original Text: 22 June 1990
Rephrase Text: June twenty-two
nineteen ninety

Original Text: at least one
Rephrase Text: at lest 1

Original Text: twenty-five
Rephrase Text: 25

Original Text: 2001
Rephrase Text: two thousand and 1

Original Text: 2 years
Changed Text:
```

A.4 Few-shot Prompt for Evaluating Factual consistency

```
Decide if the claim is faithful
with the corresponding context.
Note that Factual consistency
means all information
in the claim is supported
by the context.
Answer with 0 (consistent),
1 (no evidence), or
2 (contradiction).

Context: I burst through a set
of cabin doors, and fell to
the ground-
Claim: I burst through the
doors and fell down.
Answer: 0

Context: Fun for adults and
children.
Claim: Fun for only children.
```

Answer: 2

Context: Thebes held onto power until the 12th Dynasty, when its first king, Amenemhet I who reigned between 1980 1951 b.c. established a capital near Memphis.
Claim: The capital near Memphis lasted only half a century before its inhabitants abandoned it for the next capital.
Answer: 1

[...]