ROME: Towards Robust Metrics of Factual Consistency with Sentence-Level Contrastive Alignment and Chain-of-Thought

Anonymous submission

Abstract

While the capabilities of language models have been extensively discussed, they remain prone to hallucinations and factual inconsistencies. Specifically, despite the burgeoning interest in the application of pre-trained language models for automatic evaluation metrics, we find that these widely used models struggle with longer texts and are susceptible to various adversarial attacks. In response, we propose a sentence-level evaluation method that reflects the factuality consistency between input and output, and introduce ROME. Further, we propose a Fact Chain-of-Thought (FactCoT) to elicit LLMs to construct a robust meta-evaluation benchmark encompassing various types of errors and approximately 50k factuality-consistent datasets based on six human-annotated datasets. Integrating three contrastive objectives to bolster model robustness against adversaries, ROME is a sentence-level model that can be expanded to handle long inputs and detect outputs with factual inconsistencies. When applied to address the issue of factual inconsistencies in text summarization tasks, ROME's performance significantly surpasses existing models. It further demonstrates its generalizability to unseen tasks.

1 Introduction

Despite the readability of summaries generated by abstractive summarization models, they often present unfaithfulness issues or factual inconsistencies in text summarization and natural language generation(Maynez et al., 2020; Chen et al., 2021). Previous studies show that about 30% of summaries from leading systems exhibit faithfulness errors (Cao et al., 2018; Kryscinski et al., 2020). Thus, it is paramount to utilize reliable evaluation metrics to track and gauge the faithfulness of these systems' text generations. Traditional NLG metrics such as ROUGE (Lin, 2004) and METEOR (Lavie and Agarwal, 2007) are n-gram overlap-based and show a weak correlation with human evaluations of factual consistency. Recent metrics have been proposed that calculate semantic similarity with pre-trained models (Sai et al., 2021; Zhang et al., 2020), or incorporate auxiliary tasks like textual entailment (NLI) (Mishra et al., 2021) and question answering (QA) (Durmus et al., 2020) for better faithfulness evaluation. Many prevalent model-based evaluation metrics have limitations in processing long texts and are vulnerable to adversarial attacks, compromising their reliability (Sai et al., 2021; Pagnoni et al., 2021a). To overcome these limitations, we propose a new robust evaluation metric ROME suitable for long text summarization.

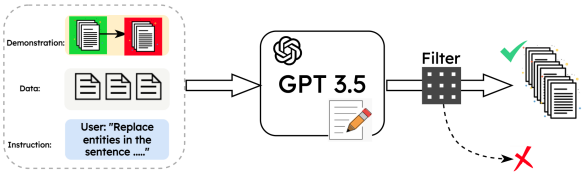

Figure 1: Data generation pipeline. Specifically, we input demonstration of transformation, input data, and Fact Chain-of-Thought prompts into GPT-3.5 to generate new transformations, which are then filtered using a consistency checker to ensure their validity. The resulting data is utilized as the final adversarial training dataset.

In this paper, we demonstrate the performance of our model, ROME, by establishing a robust meta-evaluation benchmark specifically designed for assessing factual consistency. Traditional benchmarks reliant on human annotation are not scalable for large-scale data, and using tools such as Stanford CoreNLP toolkit (Manning et al., 2014) and NLTK WordNet (Bird and Loper, 2004) for data augmentation can be easily exploited by models. These tools fail to provide a comprehensive coverage of all error types and fall short in delivering effective generalization checks. To counter these limitations, we employ synthetic text generated via the GPT-3 API (Brown et al., 2020), and use the

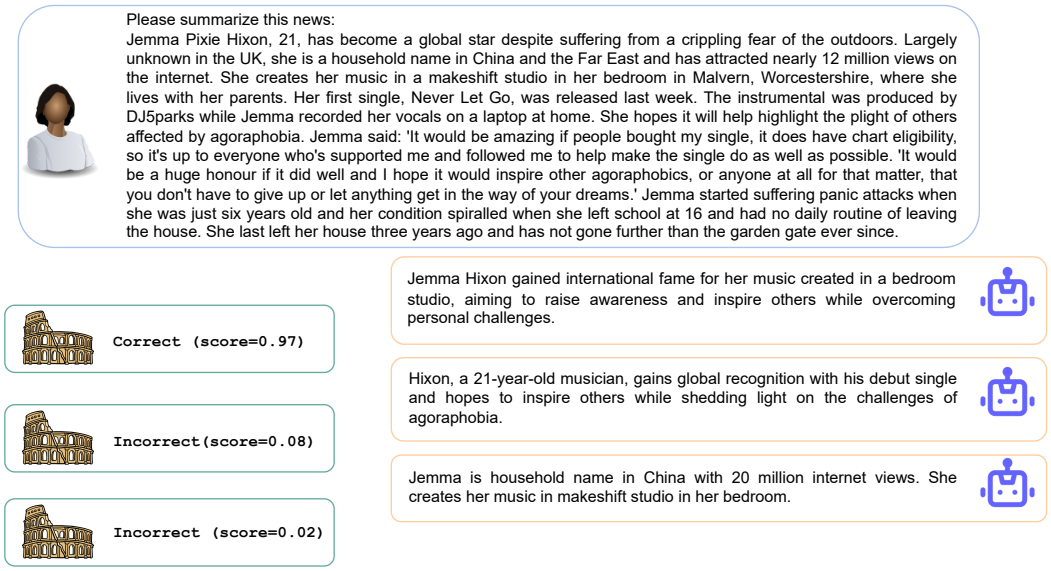

Figure 2: ROME assessing factual consistency between input and output: Three outputs generated by different models. The first output is accurate, the second contains factual inconsistency related to character gender, and the third displays numerical factual inconsistency.

Chain-of-Thought (CoT) method to create accurate variations and ensure the precision of adversarial attacks. Our rigorous manual inspection affirms that our extensive meta-evaluation benchmark exhibits an impressively low error rate of 0%. Moreover, our adversarial meta-evaluation framework offers granular evaluation of factual metrics based on diverse error types. We find that the performance of existing state-of-the-art metrics declines in most cases, revealing that many metrics lack robustness and reliability. Our approach thereby ensures non-exposure of attack types used during training in the testing phase, adding a layer of security.

Dataset	# Valid	# Test	% Positive
CGS	1281	400	49.7
XSF	996	996	9.4
Polytope	634	634	87.2
FactCC	931	503	85.8
SummEval	850	850	90.6
FRANK	671	1575	33.2
Ours	5000	5000	50

Table 1: Statistics of the six previous datasets and our benchmark.

Our ROME framework leverages two types of base models, each built upon ELECTRA (Clark et al., 2020) and LLaMA (Large Language Models as Meta AI) (Touvron et al., 2023), respectively. We introduce an innovative two-stage model training process that takes into account the scale and quality of data from various sources. In addition to the standard binary classification target, we employ supervised contrastive loss (Khosla et al., 2020) to magnify the differences between similar statements bearing different correctness labels. Moreover, we propose a method for augmenting training data to cover various error types using CoT.

Our contributions in this work are as follows:

- We formulate a new problem: the robustness of factual consistency evaluation. By leveraging GPT-3.5 and the Chain of Thoughts (CoT) approach, we generate diverse synthetic benchmarks and training data based on different transformation types while keeping the attack types undisclosed.

- We implement an effective sentence-matching scheme that retrieves relevant pairs from source and target data. This is followed by a contrastive training phase, contributing to the simplicity and efficiency of our approach.

- We demonstrate an enhancement in our model's robustness, increased generalization capabilities on diagnostic benchmarks, and significant improvements in robustness across a range of attack scenarios, surpassing previous works. Furthermore, we provide evidence of its effectiveness in real-world applications.

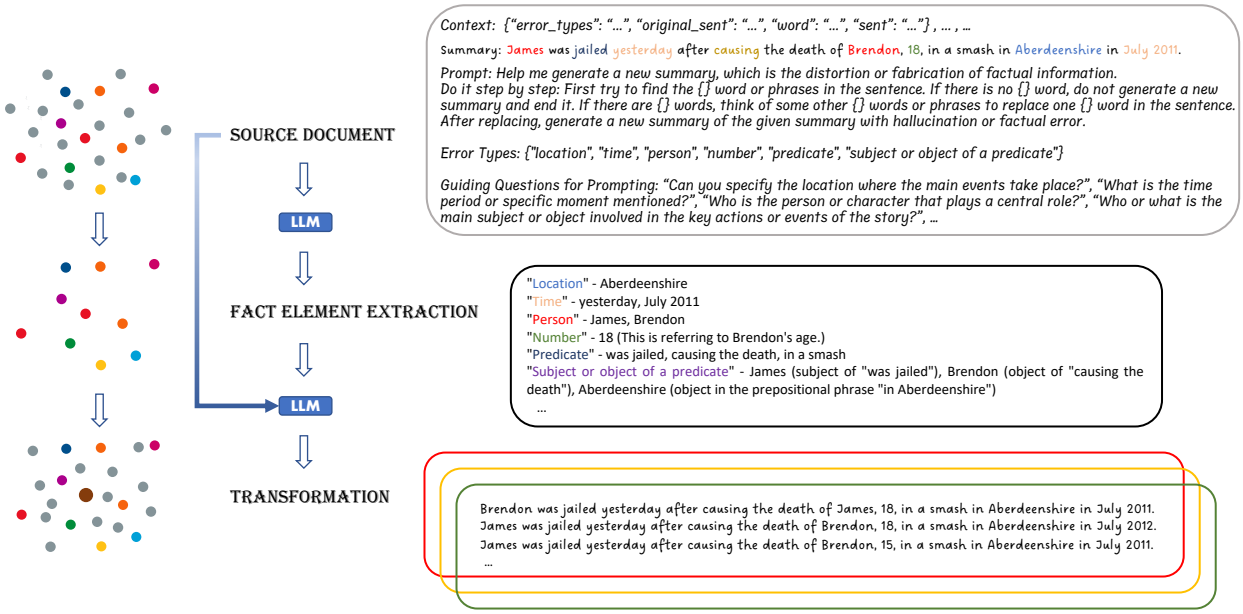

Figure 3: Pipeline and example of our Fact Chain-of-Thought.

Negative Transformation	Explanation
Location	*Incorrect or misleading geographical information is introduced.*
Time	*The temporal context or time-related details are inaccurately depicted.*
Person	*The wrong person or entity is attributed to an event or action.*
Number	*Numerical information, such as quantities or measurements, are misrepresented.*
Predicate	*Actions or events are described inaccurately or in a misleading manner.*
Pronoun	*Incorrect or misleading usage of pronouns, possibly leading to confusion about entities or actions.*
Negation	*The original meaning is distorted by adding or removing negation.*
Subject or object	*The subject or object associated with an event or action is incorrectly represented.*

Table 2: Explanation of Different Error Types

2 Related work

Addressing the shortcomings of traditional metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), there has been an emergence of automatic model-based metrics. In evaluating factual consistency, a popular approach is to employ question answering (QA)-based metrics, such as QAGS (Wang et al., 2020) and FEQA (Durmus et al., 2020), which involve generating and answering questions about a summary. However, while these QA-based metrics offer interpretability, they come with computational costs due to the utilization of QG and QA models. Textual-entailment-based or NLI-based metrics have also been widely adopted, proposed by (Falke et al., 2019), which consider a summary factually consistent if it is semantically entailed by the source document. These metrics, however, face challenges due to domain shifts and reliance on heuristics like lexical overlap (Mishra et al., 2021). To improve generalizability, researchers have suggested the use of long-premise NLI datasets, which led to significant improvements in factual consistency evaluation (Mishra et al., 2021). Despite these advancements, NLI-based metrics such as FactCC (Kryscinski et al., 2020) are vulnerable to adversarial attacks and suffer from a misalignment with the actual NLI task due to overly long premises.

3 Preliminaries

3.1 Task Definition and Scope

In our research, we conceptualize the evaluation of factual correctness or faithfulness in a summa-

rization system as a Natural Language Inference (NLI) task. In this context, a source document and a generated summary are dispatched as a pair to an NLI classifier, which subsequently assesses the relationship between them, specifically whether the summary is faithful or unfaithful to the original content. A summary is considered faithful if it accurately encapsulates the information of the source without introducing any factual errors.

Historically, researchers have used human annotation for meta-evaluation, where annotators categorize whether summaries contain hallucinated facts. This meta-evaluation forms the groundwork for building a benchmark to evaluate the factual consistency of different metrics. In contrast, our approach substitutes human labor with synthetic data generation, producing deterministic non-factual summaries and factual summaries.

We define **Factual Consistency** as the extent to which the summary accurately mirrors the content of the original document without incorporating any factual inaccuracies. On the other hand, **Robustness** signifies the resilience of NLI-based evaluation metrics against manipulative attacks, maintaining their accuracy even in the face of intentionally altered inputs.

Our research focuses on the robustness of summarization system evaluation metrics, particularly in terms of faithfulness, or factuality. We place particular emphasis on the metrics' resilience to alterations that modify the factual content of sentences.

3.2 Meta-evaluation Benchmark and FactCoT

Evaluating a model's ability to detect factual inconsistencies between input and output requires labeled data. Unfortunately, such labeled factual inconsistency statements do not often occur in the wild. In the past, considerable efforts have been made to manually annotate such data. However, relying on crowdsourced annotations often results in unsatisfactory quality and it's hard to amass a large-scale dataset suitable for training. Recently, some works have utilized tools like NLTK and Named Entity Recognition (NER) for automatic adversarial transformations. Yet these methods are constrained by the types of errors they can introduce, and the resulting factual errors tend to be homogeneous. This uniformity can lead to overfitting, as models may easily memorize these errors rather than genuinely learning to generalize their capabilities. To address these issues, we utilize

Type	***Prompt***
Pronoun	*Help me generate a new summary by pronoun swapping transformation. Do it step by step: First try to find the gender-specific pronouns words in the sentence. If there are no gender-specific pronouns words, do not generate a new summary and end it. If there are gender-specific pronouns words, Next, a randomly chosen pronoun was swapped with a different one from the same pronoun group to ensure syntactic correctness, i.e., a possessive pronoun could only be replaced with another possessive pronoun. After replacing, generate a new summary of the given summary with incorrect pronoun use. Don't change too much, just change the gender-specific pronouns words.*
Negation	*Help me generate a new summary by sentence negation transformation. Do it step by step: First try to find the auxiliary verbs in the sentence. If there are no auxiliary verbs, do not generate a new summary and end it. If there are auxiliary verbs, next, to switch the factual meaning, a randomly chosen auxiliary verb was replaced with its negation. Positive sentences would be negated by adding not or n't after the verb, and negative sentences would be switched by removing the negation. After replacing, generate a new summary of the given summary with sentence negation transformation. Don't change too much, just change the auxiliary verbs.*

Table 3: Prompts for Different Transformation Types

Large Language Model (LLM) to perform transformations, which allows us to construct a new robustness benchmark or a diagnostic dataset. This implies that our dataset can serve dual purposes: it can be employed as part of the test set to diagnose model performance or be integrated into the training set to enrich the data for augmentation purposes. This approach enables us to both expand the diversity and scale of our training data and enhance the evaluation of a model's robustness to diverse factual errors. These transformations are applied to existing human-annotated datasets. For the generation of synthetic training data, we feed prompts to GPT-3. Inspired by LLM's competitive zero-shot performance, especially in-context learning and Chain of Thought (CoT), we create a Fact Chain of Thought (FactCoT) to guide LLM to gradually generate summaries with certain error types. Figure 3 provides examples of these prompts. Here the example presented includes error types such as "location", "time", "person", "number", "predicate", and "subject or object of a predicate". However, error types like "pronoun" and "negation" are more complex. After experimental attempts, we selected more intricate prompts, which can be seen in Ta-

ble 3. We generated 50k data, randomly selecting 500 for manual evaluation. No error or noise was discovered, indicating a remarkable quality, which significantly exceeds that of data annotated manually or using NER tools for augmentation.

It's noteworthy that we simultaneously utilize the XSum and CNN/DM datasets in our research. The XSum dataset exhibits a considerably larger quantity of unfaithful summaries compared to CNN/DM (Pagnoni et al., 2021b; Li et al., 2022). This is attributed to the heuristic collection method of the XSum dataset, where introductory sentences and article beginnings are used as reference summaries. As a result, these reference summaries typically contain hallucinations (Narayan et al., 2018a).

We incorporate two types of transformations in our process. The first one modifies human-crafted correct summaries to augment the dataset, thus balancing the distribution of positive and negative labels. The second type of transformation introduces factual errors into the summary. These transformations allow us to selectively introduce errors of different types, including those related to "location", "time", "person", "number", "predicate", "pronoun", "negation" and "subject or object of a predicate", as shown in Table 2. Furthermore, the positive transformations encompass the following types: "person", "predicate", "entity", and "subject or object of a predicate". These types are similar to the negative transformation but do not change the factual consistency label of the summary.

4 Methodology

As LLM does not explicitly learn factual inconsistency during pre-training, and directly fine-tuning them using labeled data might lead to overfitting due to excessive memorization. To facilitate the model in discerning the distinction between factual inconsistencies in the input and output, we have devised an improved learning method that encourages the model to better understand and capture this alignment.

4.1 Formulation

Consider a document D with sentences $d_1, d_2, \ldots, d_N$ and a summary S composed of sentences $s_1, s_2, \ldots, s_M$. Using these inputs, our method ROME outputs a score s in the range of 0 to 1. ROME utilizes transformer-based models, initially, we have the final hidden state related to the End of Sentence (EOS) token to get the input representation h. EOS is chosen because it can effectively encode the complete input in both bidirectional models like ELECTRA and decoder-only models like LLaMA. ELECTRA applies a unique pre-training task known as Replaced Token Detection (RTD), which allows it to train bidirectionally and learn from all input positions.

4.2 ROME Metrics

We leverage three loss functions in our approach. The first function encourages the model to assign higher scores to correct sentences, treating it akin to a binary classification task. However, considering that in generated data, incorrect summaries often outnumber the correct ones, we normalize this loss by the number of summaries with the same label in the same batch.

The second function prompts the model to assign higher scores to correct propositions compared to incorrect ones. We desire the model to exhibit robustness towards subtle differences in the summaries, as factual inaccuracies can often be minor and difficult to spot. Ideally, the model should be able to discern opposing factual consistency labels for a group of summaries that may seem similar in their surface form but contain subtle factual inaccuracies pertaining to the input. Although these summaries might be semantically similar from the perspective of an NLI model, their subtle factual discrepancies relative to the input make them distinct. We treat summaries created through different transformations for the same input as a multi-class classification problem and maximize the log-likelihood of a single correct summary in the statement group after passing logits through softmax.

$$
\begin{aligned}
\mathcal{L} = \alpha \Bigg[& \frac{1}{B_G} \sum_{j=1}^{B_G} \Bigg\{ \sum_{y \in \{0,1\}} \\
& \Bigg[\frac{1}{\sum_{c=1}^{C_j} I[y_{jc} = y]} \sum_{c=1}^{C_j} I[y_{jc} = y] \\
& (-y_i \log s(x_{jc}) - (1 - y_i) \log(1 - s(x_{jc}))) \Bigg] \Bigg\} \Bigg] \\
& + \beta \left[\frac{1}{B_G} \sum_{j=1}^{B_G} \left(- \log \frac{\exp z(x_{j*})}{\sum_{c=1}^{C_j} \exp z(x_{jc})} \right) \right] \\
& + \gamma \Bigg[\frac{1}{B_S} \sum_{i=1}^{B_S} \Bigg(- \log \\
& \frac{\sum_{k \in \mathcal{P}(i)} \exp\left(\frac{\cos(\mathbf{h}(x_i), \mathbf{h}(x_k))}{\tau} \right)}{\sum_{k \in \mathcal{P}(i) \cup \mathcal{N}(i)} \exp\left(\frac{\cos(\mathbf{h}(x_i), \mathbf{h}(x_k))}{\tau} \right)} \Bigg) \Bigg]
\end{aligned}
\tag{1}
$$

Further, with the third function, we aim to learn

a representation for the learned summary. We try to learn the similarity within a positive sample batch while distancing it from other negative samples. This function is a classic instantiation of a contrastive loss. In the equation for $\mathcal{L}$, the total loss function we aim to minimize consists of three major parts, each component weighted by α, β, and γ, respectively.

The first term sums over instances in a binary group B_G, with j denoting each instance and c indexing over classes C_j. y_{jc} is the ground truth label, with $y \in 0, 1$, and $I[y_{jc} = y]$ is an indicator function that returns 1 if $y_{jc} = y$ and 0 otherwise. The binary loss is computed using $s(x_{jc})$, which is a sigmoid function of input x_{jc}.

The second loss component is computed using the softmax of $z(x_{jc})$ over all the classes c for each instance j in the binary group B_G. The softmax function is used to convert raw model outputs to probabilities that sum to 1.

The last contrastive loss component is calculated over a different batch B_S, with i indexing each instance in this batch. The cosine similarity between the representations $\mathbf{h}(x_i)$ and $\mathbf{h}(x_k)$ of each instance and all other instances k in a set of positives $\mathcal{P}(i)$ or negatives $\mathcal{N}(i)$ is computed, and these similarities are passed through an exponential function. The temperature hyperparameter τ is used to control the concentration of the distribution, i.e., the sharpness of the softmax function.

Furthermore, we aim to improve the process of evaluating the factual consistency between a document and its summary. Conventionally, these evaluations have been carried out on a document-wide scale, producing a singular score that indicates the factual alignment between the entirety of the document and its summary. However, this approach may not accurately capture factual inconsistencies that occur within distinct parts of the document.

To address this issue, we introduce a novel scoring method. We employ a matching model to discern the top three sentences in the document that holds the most relevance to the summary, treating these as evidential segments. By focusing our evaluation on these selected excerpts and minimizing the potential noise from the broader document, we specifically target factual consistency of key events.

The improved scoring methodology can be mathematically expressed as follows:

$$score_{fact} = \frac{1}{j}\sum_{k=1}^{j} \max_{i \leq k} \left(\frac{1}{i} \sum_{l=1}^{i} \text{score}(d_l, s_k) \right) \quad (2)$$

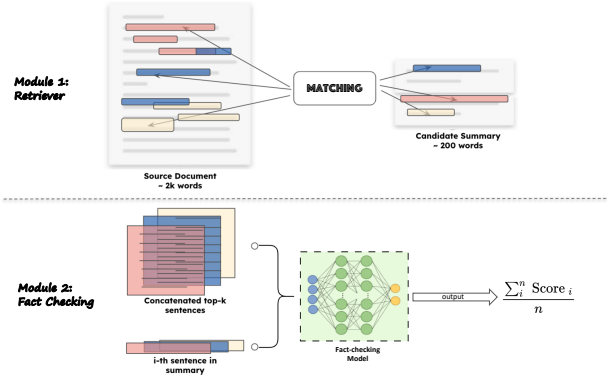

Figure 4: Illustration of the two-module NLI-based evaluation metric.

In this formulation, the document D is divided into i segments up to the jth sentence of the summary s. The score function signifies the factual consistency between a chosen document segment d_l and the jth sentence of the summary s_k. By taking the weighted average of these scores, we obtain a measure that more precisely reflects the factual concordance of each key section of the document with the summary. This sentence-level technique provides a more nuanced and precise evaluation of factual consistency compared to conventional document-level scoring methods.

ROME employs a two-module design, as illustrated in Figure 4, the Sentence Matching (Retriever) module utilizes a neural matching model to identify the Top-k relevant sentences in the source document that correspond to each sentence in the summary. Next, the Fact-checking module evaluates the accuracy of each summary sentence against the concatenated top-k document sentences retrieved. We use the mean aggregation of all subcomponents (Doc$_{topk}$-sentence pairs) for an input example. We argue that sentence-level evaluation approach has the potential to improve evaluation accuracy, and offers better explainability, which provides an avenue for future research on building trustworthy explainable evaluation metrics.

5 Results

5.1 Setting

As it is not feasible to produce large-scale human-constructed benchmarks (Xie et al., 2021; Falke et al., 2019; Kryscinski et al., 2020), as described before we propose FactCoT, considered as adversarial attacks, to automatically construct diagnostic benchmarks. We select source data and annota-

tions, which include human-written summaries and labels, from previous works such as the CNN/DM dataset (Nallapati et al., 2016) and XSUM dataset (Narayan et al., 2018b).

We select 6 baseline metrics and test them on our curated diagnostic benchmarks. ROUGE (Lin, 2004) is a traditional text generation evaluation metric based on n-gram overlap, which cannot assess the faithfulness of text generation. BERTScore (Zhang et al., 2020)is an encoder-based automatic evaluation metric. Unlike ROUGE which computes token-level syntactical similarities, BERTScore focuses on computing semantic similarity between tokens of reference and hypothesis. Besides, we use 3 NLI-based metrics. BERT-base (Devlin et al., 2019) and RoBERTa-large (Liu et al., 2019) are variants that are fine-tuned on the Multi-Genre NLI (MNLI) corpus (Williams et al., 2018). FactCC (Kryscinski et al., 2020) is a recent NLI-based metric, where it constructed positive entailment samples by different heuristic methods for weakly supervised training. Finally, we include a recent popular QA-based metric FEQA (Durmus et al., 2020) for comprehensiveness. We present our baseline results in the following tables, which include the original scores and the absolute change in scores under our test sets. Our results demonstrate that all baseline metrics suffer to varying degrees. We also note trivial gains in the correlation score of FactCC (NLI) (Kryscinski et al., 2020) under the benchmark derived from the XSUM dataset (Narayan et al., 2018b).

Due to the unbalanced distribution of positive and negative examples in the test datasets, balanced accuracy is chosen as a metric: $bACC = \frac{1}{2} * \left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$, accounting for the performance of both positive and negative classes by calculating the average of the true positive rate (recall) and the true negative rate (specificity) (Brodersen et al., 2010).

6 Experiments

6.1 Data

We generate synthetic $(source, summary, label)$ pairs for training data by selecting samples from the held-out fraction during the benchmark test-set construction. In essence, we utilize examples that were not included in the creation of the diagnostic benchmark to produce diverse synthetic training data. We use the GPT-3.5 API (175B; `davinci`) (Brown et al., 2020) to create our training samples through in-context learning. This approach involves conditioning the model on a few examples and instructional prompts, as described in (Dong et al., 2023). This process consists of two steps, as illustrated in Figure 3. We input into the system 1) some original data examples, i.e., a (document, summary) pair, 2) 10 demonstrations of original and transformed sentences of a specific error type, and 3) a set of user messages to instruct the system to generate synthetic data based on the error types. Next, we filter out invalid generations using the code example in Appendix C, as some sentences cannot undergo the expected transformation. Our goal is to generate a total of 50k document sentences, resulting in 45k examples for training our metric. We ensure that the training data is diverse and reliable while also protecting the test set from exposure to specific attack types.

6.2 Baselines

We select 6 baseline metrics and test them on our curated diagnostic benchmarks. ROUGE (Lin, 2004) is a traditional text generation evaluation metric based on n-gram overlap, which cannot assess the faithfulness of text generation. BERTScore (Zhang et al., 2020)is an encoder-based automatic evaluation metric. ELECTRA and RoBERTa-large (Liu et al., 2019) are variants that are fine-tuned on the Multi-Genre NLI (MNLI) corpus (Williams et al., 2018). FactCC (Kryscinski et al., 2020) is a recent NLI-based metric, where it constructed positive entailment samples by different heuristic methods for weakly supervised training. Finally, we include a recent popular QA-based metric FEQA (Durmus et al., 2020) for comprehensiveness. We present our baseline results in the following tables, which include the original scores and the absolute change in scores under our test sets. Our results demonstrate that all baseline metrics suffer to large degrees. As we will see in Table **??**, ROME-ELECTRA has a better performance than ROME-LLaMA.

6.3 Training Details

For LLaMA, we use the pre-trained LLaMA-7B. We use a learning rate of e-5 for ELECTRA and 2e-6 for LLaMA, a batch size of 32 examples, and the default random seed 42. The best model checkpoints were selected based on their performance on the validation set, while the final model performance was assessed on the test set. Our experiments were conducted using eight NVIDIA

Metric \Benchmark	Original		Sub or obj		Number		Location	
	XSUM	CNN/DM	XSUM	CNN/DM	XSUM	CNN/DM	XSUM	CNN/DM
ROUGE-2	21.06	73.93	**- 1.48**	-40.43	**- 0.35**	-42.32	- 2.31	-34.55
BERTScore	19.08	71.77	- 35.10	-41.35	- 26.08	-23.10	- 42.45	-29.68
ELECTRA	61.92	79.92	- 21.45	-45.98	- 24.13	-56.31	-25.23	-43.71
RoBERTa-MNLI	41.12	83.30	- 21.41	-66.13	- 6.43	-52.34	- 14.19	-48.33
FEQA	91.59	77.93	**+ 1.45**	-48.64	- 3.35	-35.84	- 5.30	-27.34
FactCC	77.32	86.08	- 11.34	-73.29	**+ 3.41**	-30.20	**+ 1.91**	-5.39
ROME-LLaMA7B$_P$	57.72	61.26	-18.98	-17.51	-18.45	-12.58	-13.01	-10.91
ROME-LLaMA7B$_S$	62.54	79.63	-11.29	-12.24	-12.36	-11.01	-5.78	-4.89
ROME-ELECTRA$_P$	61.92	79.92	-9.37	-11.56	-8.89	-13.72	-5.54	-5.21
ROME-ELECTRA$_S$	71.31	88.87	-4.88	**-8.62**	**-5.07**	**-2.32**	**-2.89**	**-0.78**

Table 4: Absolute change in accuracy scores under 3 transformations based on Faith(XSUM) and FactCC(CNN/DM) datasets by 6 baseline metrics.

GeForce RTX 3090 GPUs.

6.4 Results

As indicated in the tables, our model significantly improves upon the baselines across all attack scenarios. This finding illustrates that our approach increases the model's robustness when faced with these various types of adversarial attacks. The improvements observed across various attack types highlight the effectiveness of our approach in enhancing the model's resistance to a range of adversarial attacks, ultimately leading to better factual error classification in text summarization evaluation.

And our sentence-level approach outperforms the document-level approach for both datasets (and annotations) and for all diagnostic datasets constructed under different types of attacks. For example, intuitively, we note that some number changes can be more easily detected by analyzing the most relevant sentences instead of the entire document. In contrast, the doc-level approach considers the whole document as a single unit, potentially confluent with multiple numerical mentions.

In addition, we also found 5 examples of factual inconsistencies produced by ChatGPT on the Internet, and our model can make accurate judgments, so We find ROME can be used to scrutinize factual inconsistencies in outputs produced by models like ChatGPT in real- world scenarios.

6.5 Generalizability and Practicality

Regarding ROUGE-2, despite minor performance variations on adversarial data in certain categories, the overall score is particularly poor, rendering it impractical. As for FEQA, while there's minimal performance fluctuation on XSUM, it is remarkably unrobust on CNN/DM. This could potentially be attributed to the data used during its training phase. Overall, our model exhibits robust generalizability and effective fact consistency check results.

7 Conclusion

In this paper, we introduce ROME, a novel model designed to assess the factual consistency between inputs and outputs. We put forth the concept of robustness in factual consistency checking, identifying an overlooked challenge in current state-of-the-art models. Despite their impressive performance, we found these models to demonstrate significant performance degradation under our robustness benchmark. This is likely due to the identical distribution of their training and testing datasets. To overcome these limitations, we employed LLaMA and ELECTRA to train our ROME model, striving for a model less susceptible to such distributional shifts. Our findings reveal that ROME exhibits minimal average disparity between scores obtained on the diagnostic test and original scores, outperforming previous models in robustness. Additionally, we observed that implementing sentence-level checks led to a significant boost in performance. Integrating LLM into data augmentation offers a promising avenue for advancing the robustness of factual consistency checks.

8 Limitation

We conduct an analysis of the false output by our models, and we admit the explainability limitations in our project. We provide some examples of errors in Appendix C. Major patterns are Missing Details, Co-reference Failure, Negation, and Extrapolation of Training data. We suspect the matching scheme based on nerual matching model may cause the

model to miss relevant information that falls outside the matched sentences to cause co-reference failure. Also, even when details are present in the matched sentences, the model cannot incorporate them into its prediction. This suggests that the adversarial synthetic data may still not adequately represent the diverse examples. Besides, we realized that our generated adversarial training data could reflect the certain distribution of the test data, but the generalization gap still exists.

References

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. 2010. The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*, pages 3121–3124.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact-aware neural abstractive summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5935–5941, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

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.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

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.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. *arXiv preprint arXiv:2203.05227*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

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.

Anshuman Mishra, Dhruvesh Patel, Aparna Vijayakumar, Xiang Lorraine Li, Pavan Kapanipathi, and Kartik Talamadupula. 2021. Looking beyond sentence-level natural language inference for question answering and text summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1322–1336, Online. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018a. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018b. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021a. 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.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021b. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. *arXiv preprint arXiv:2104.13346*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ananya B. Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M. Khapra. 2021. Perturbation checklists for evaluating nlg evaluation metrics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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.

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. Association for Computational Linguistics.

Yuexiang Xie, Fei Sun, Yang Deng, Yaliang Li, and Bolin Ding. 2021. Factual consistency evaluation for text summarization via counterfactual estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 100–110, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Example Sentences
Original: Albert Einstein, a renowned theoretical physicist, is best known for his theory of relativity and the famous equation E=mc^2. *Transformed:* Einstein, a renowned theoretical physicist, is best known for his theory of relativity and the famous equation E=mc^2. *Label:* IN USE *Tag:* person
Original: In 1986, the Chernobyl disaster, a catastrophic nuclear accident, took place in the Soviet Union, releasing large amounts of radioactive materials into the environment and causing significant long-term health and environmental consequences. *Transformed:* The Chernobyl disaster, a catastrophic nuclear accident, occurred in the Soviet Union, resulting in the release of large amounts of radioactive materials into the environment and leading to significant long-term health and environmental consequences. *Label:* IN USE *Tag:* circumstance
Original: The Industrial Revolution, which occurred between the late 18th and early 19th centuries, was a period of rapid industrialization, urbanization, and technological advancements that drastically transformed society and the global economy *Transformed:* The Industrial Revolution, taking place between the late 18th and early 19th centuries, was a time of fast industrialization, urbanization, and technological progress that significantly altered both society and the worldwide economy. *Label:* IN USE *Tag:* complex

Table 5: Example Summary Transformation.

A Examples of Transformations

Positive examples: see Table 5. Negative examples: see Table 6.

B Example of system prompts

Example Sentences
Original: James Watson was jailed yesterday after causing the death of Brendon Main, 18, in a smash in Aberdeenshire in July 2011. *Transformed:* James Watson was jailed yesterday after causing the death of Brendon Main, 18, in a smash in Madrid in July 2011. *Label:* IN USE *Tag:* location
Original: Brown was unarmed when he was fatally shot by a white police officer in a St. Louis suburb in August 2014. *Transformed:* Brown was unarmed when she was fatally shot by a white police officer in a St. Louis suburb in August 2014. *Label:* IN USE *Tag:* pronoun
Original: Snow was predicted later in the weekend for Atlanta and areas even further south. *Transformed:* Snow wasn't predicted later in the weekend for Atlanta and areas even further south. *Label:* IN USE *Tag:* negation

Table 6: Example Summary Transformation.

Property	**Description**
Requirement	Analyze a given sentence's grammatical structure and modify it in various ways
Requirement_0	Ensure the new sentence has a different meaning
Iterate_way	Replace entities in the sentence to create new sentences
Ret_format	Return sentence structure
Iterate_ret_format	Return sentences with replaced entities

Table 7: System properties and descriptions for GPT-3 API input

Category	Description
User Input (Case 1)	The 25-year-old will face Marco Fu in second round of World Championship.
Sentence Structure (Case 1)	The 25-year-old [subject] will [modal verb] face [verb] Marco Fu [direct object] in second round [prepositional phrase] of World Championship [object of preposition].
Subject	The one wearing glasses will face Marco Fu in second round of World Championship.
Verb	The 25-year-old won't kiss Marco Fu in second round of World Championship.
Object	The 25-year-old will face Ronnie O'Sullivan in second round of Speech competition.
Preposition	The 25-year-old will face Marco Fu after second round of World Championship.
Clause	not_valid
Number	The 18-year-old will face Marco Fu in first round of World Championship.
Sequence	The 25-year-old has already faced Marco Fu in second round of World Championship.
Name	The 25-year-old will face Taylor Swift in second round of World Championship.
Date & Time	not_valid
Location	not_valid
User Input (Case 2)	T-shirts joking about stalking are sold online in various US and UK stores.
Sentence Structure (Case 2)	T-shirts [subject] joking [gerund] about stalking [prepositional phrase] are sold [verb] online [adverb] in various US and UK stores [prepositional phrase].
Subject	Shoes joking stalking are sold online in various US and UK stores.
Verb	T-shirts condemning stalking are discussed online in various US and UK stores.
Object	not_valid
Preposition	not_valid
Clause	not_valid
Number	not_valid
Sequence	not_valid
Name	not_valid
Date & Time	not_valid
Location	T-shirts joking about stalking are sold online in various China and Sweden stores.

Table 8: Examples of generated data based on the prompts

C Additional Details

[language=Python] def is_factual_gpt35(sent1: str, sent2: str) -> bool: messages=["role": "system", "content": "Determine whether the two sentences are in factual consistency. Briefly explain step-by-step how the second sentence changes or retains the factuality of the first sentence. If they are factually entailed, output a final answer ÿes̋, otherwise, answer ñő. ", "role": "user", "content": "".format(sent1, sent2)] expected output: "no" or "False" response = openai.ChatCompletion.create(model="gpt-3.5-turbo", or use GPT-4 for future work. messages = messages, other parameters) Return True if final response includes "yes". False otherwise.

: Filtering process in synthetic data generation

Document Text	Claim	Label	Predicted Label	Error Type
A new study has revealed that coffee drinkers are less likely to develop liver cancer...	Drinking coffee decreases the likelihood of getting liver cancer.	YES	NO	Misunderstanding of Negation
With a team of more than 130 meteorologists and 240 broadcasters...	The Weather Channel has the largest team of meteorologists and broadcasters.	NO	YES	Ambiguity in Pronoun Reference
The International Space Station was visited by two astronauts...	Two astronauts visited the International Space Station.	YES	NO	Confusion of Active and Passive Voice
According to a recent survey, more than 70% of people...	The majority of people support the use of renewable energy.	YES	NO	Misinterpretation of Quantifiers
Leicester's Premier League forecast has looked gloomy for some considerable time...	Vardy scored an injury-time winner to improve his side's slim chance of survival.	YES	NO	Missing Details, Incomplete Information
The new iPhone is expected to be released in September with a larger screen and better battery life.	The new iPhone will have a smaller screen and shorter battery life.	NO	YES	Negation Error, Misunderstanding of Context
More than 20,000 people attended the concert...	Fewer than 20,000 people attended the concert.	NO	YES	Misunderstanding of Negation
The study found that the new drug was effective in treating the symptoms of the disease...	The study found that the new drug was ineffective in treating the symptoms of the disease.	NO	YES	Negation Error, Misunderstanding of Context
The company announced that it will be launching a new product line...	The company announced that it will be discontinuing its product line.	NO	YES	Negation Error, Misunderstanding of Context
A recent study showed that people who eat more vegetables have a lower risk of heart disease...	Eating more vegetables does not lower the risk of heart disease.	NO	YES	Misunderstanding of Negation, Misinterpretation of Quantifiers, Influence of Extrapolation of Data

Table 9: Ten examples of errors by our system from the two datasets.For formatting reasons, we use "YES" representing "CORRECT" and "NO" representing "INCORRECT".