AMRFact: Enhancing Summarization Factuality Evaluation with AMR-Driven Negative Samples Generation

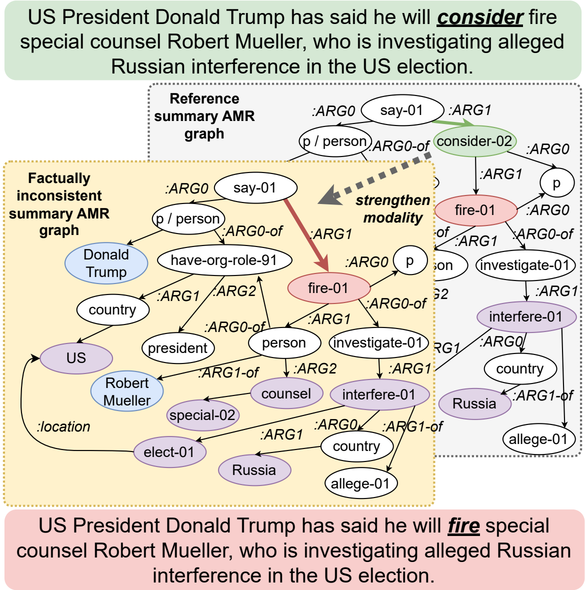
Anonymous ACL submission

Abstract

Ensuring factual consistency is crucial for natural language generation tasks, particularly in abstractive summarization, where preserving the integrity of information is paramount. Prior works on evaluating factual consistency of summarization often take the entailment-based approaches that first generate perturbed (factual inconsistent) summaries and then train a classifier on the generated data to detect the factually inconsistencies during testing time. However, the perturbed summaries produced by these approaches are either of *low coherence* or *lack error-type coverage*. To address these issues, we propose AMRFact, a framework that generates perturbed summaries using Abstract Meaning Representations (AMRs). Our approach parses factually consistent summaries into AMR graphs and injects controlled factual inconsistencies to create negative examples, allowing for coherent factually inconsistent summaries to be generated with high error-type coverage. Additionally, we present a data selection module NegFilter based on natural language inference and BARTScore to ensure the quality of the generated negative samples. Experimental results demonstrate our approach significantly outperforms previous systems on the AggreFact-FtSota dataset, showcasing its efficacy in evaluating factuality of abstractive summarization.

1 Introduction

Recent advances in text summarization, driven by the pre-trained language models, have enabled the generation of coherent abstractive summaries (Zhang et al., 2019; Raffel et al., 2019; Lewis et al., 2020). However, studies have shown that these generated summaries can be factually inconsistent with the source documents (Goodrich et al., 2019; Kryscinski et al., 2020; Pagnoni et al., 2021). This inconsistency between the generated summary and the factual information in the source document necessitates the need for assessing the *factuality* or factual consistency of the summaries.



US President Donald Trump has said he will ***consider*** fire special counsel Robert Mueller, who is investigating alleged Russian interference in the US election.

US President Donald Trump has said he will ***fire*** special counsel Robert Mueller, who is investigating alleged Russian interference in the US election.

Figure 1: Example of a reference (green) and a generated factually inconsistent summary (red) from the AMRFact dataset. Given a reference summary, we convert the text into an AMR graph (grey) and then remove " consider-02" to generate a factually inconsistent summary AMR graph (yellow). This perturbed summary strengthens the modality in the reference summary, resulting in factual inconsistency. The reference and perturbed summaries will be used as positive and negative examples, respectively.

Recent work formulated factual inconsistency detection as an entailment recognition task, predicting whether the source document entails a summary. The prevalent approach for developing an entailment-based factual consistency metric involves generating synthetic data and training a classifier on it. Kryscinski et al. (2020) treats reference summaries from CNN/DM (Hermann et al., 2015) as positive examples and applies entity-centric perturbations, such as entity replacement, to the reference summaries to create negative samples (*i.e.*, summaries factually inconsistent with the source document). However, such an approach often produces sentences with poor coherence caused by the string replacement operations. Goyal and Durrett (2021) and Utama et al. (2022) address the

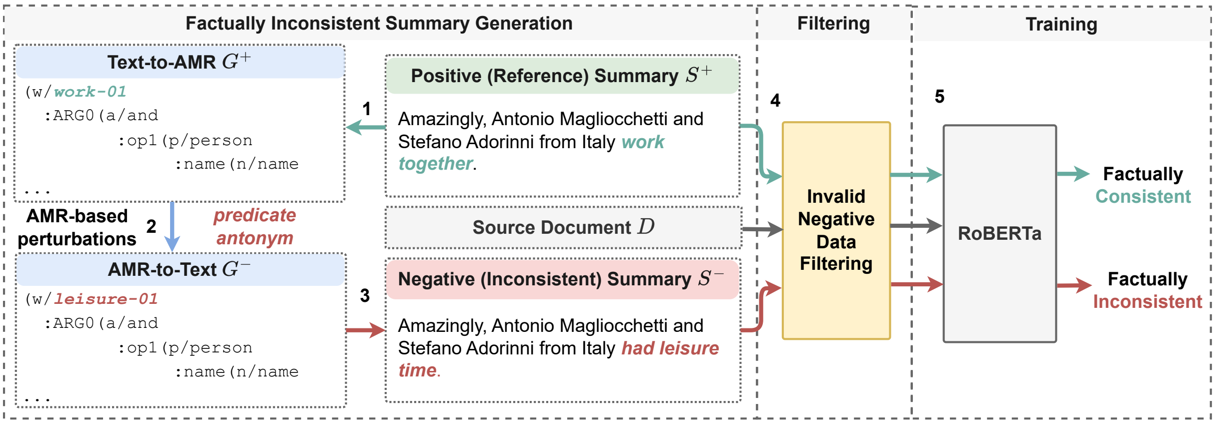

Factually Inconsistent Summary Generation
Filtering
Training
Text-to-AMR G^+
(w/work-01
:ARG0 (a/and
:op1 (p/person
:name (n/name
...
1
Positive (Reference) Summary S^+
Amazingly, Antonio Magliocchetti and Stefano Adorinni from Italy work together.
AMR-based perturbations
2
predicate antonym
Source Document D
AMR-to-Text G^-
(w/leisure-01
:ARG0 (a/and
:op1 (p/person
:name (n/name
...
3
Negative (Inconsistent) Summary S^-
Amazingly, Antonio Magliocchetti and Stefano Adorinni from Italy had leisure time.
4
Invalid Negative Data Filtering
5
RoBERTa
Factually Consistent
Factually Inconsistent

Figure 2: Overview of AMRFACT training phase: (1) The generation module first converts the reference summaries into AMR graphs. (2) These graphs are then manipulated to include common factual errors shown in current summarization systems, creating factually inconsistent AMR graphs. (3) These manipulated graphs are back-translated into text summaries, serving as negative examples for training a text-based factuality evaluator. (4) A selection module, using NLI and BARTSCORE, filters out low-quality negative examples. (5) Finally, we fine-tune a ROBERTA-based model with this data to act as the *evaluation metric*, assessing factuality by comparing the original document (premise) with the summary (hypothesis) and measuring the probability of entailment.

coherence issue by training a paraphrasing model and a text-infilling model to produce negative data. Yet, these approaches cannot ensure the production of specific types of factual errors illustrated in Pagnoni et al. (2021). As a result, models trained on data produced by these methods may not reliably detect particular types of factual inconsistency in summaries. We refer to this limitation as a lack of "*error-type coverage*". Besides, all prior methods lack a verification step to *validate* the quality of the generated data, potentially leading to diminished performance in detecting factual inconsistency.

Motivated by these challenges, we propose AMRFACT, a framework that generates **coherent negative examples** with **high error-type coverage**. Our data generation module leverages Abstract Meaning Representations (AMRs) (Banarescu et al., 2013) to introduce semantic-level perturbations for creating negative examples, enabling us to generate more coherent summaries without compromising the error-type coverage. AMRs are intended to capture the meaning of a sentence by abstracting away irrelevant syntactic features. This feature allows us to maintain coherence while precisely tailoring our negative examples to target specific factual error types. As shown in prior research (Lee et al., 2021), AMR's *controllability* empowers us to easily shape the distribution of negatives, making our approach highly adaptable and effective in addressing the identified challenges.

In detail, AMRFACT starts with parsing factually consistent summaries into semantic AMR graphs, and then injects factual inconsistencies (errors) that are *commonly* observed in state-of-the-art summarization systems (Pagnoni et al., 2021) into the AMR graphs (Ghazarian et al., 2022; Ribeiro et al., 2022). These perturbed AMR graphs are subsequently translated back into text summaries to serve as our negative examples using a controllable generation model, which ensures that the generated summaries retain a *natural* and *coherent* narrative flow. Then, we devise a novel selection module NEGFILTER to exclude invalid negative samples from our training data. A *valid* negative summary must satisfy two criteria: (1) it must not be directly inferable from the original summary and (2) it should not stray significantly from the main topic of the document. We employ sentence-level NLI and BARTSCORE to ensure compliance with each criterion, respectively. Finally, a ROBERTA (Liu et al., 2019) model is trained on the created dataset along to distinguish factually consistent and inconsistent summaries as the evaluation metric. Figure 2 offers an overview of our proposed framework. The highlights of our contributions include:

- We propose AMRFACT that uses AMR-based perturbations to generate factually inconsistent summaries, which allows for more coherent generation with high error-type coverage.

- We devise a data validation module to filter out invalid negative summaries to enhance the quality of generated data.

- Our approach achieves state-of-the-art performance on the AGGREFACT-FTSOTA dataset.

Source document fragment: (CNN) -- Martina Hingis and Anna Kournikova will team up again to play at this year's Wimbledon championships. [...] Swiss star Hingis won five grand slam singles crowns and nine in doubles during a glittering career which ended under a cloud in 2007, when she was suspended for two years for testing positive for cocaine at Wimbledon. [...]
Reference summary: Hingis has ended a two-year ban after testing positive for cocaine at 2007 Wimbledon.

Reference summary AMR graph

```
(z1/end-01
   :ARG0(z2/person
       :name(z3/name
       :op1 "Hingis"))
   :ARG1(z4/ban-01
       :ARG2 z2
       :duration(z5/temporal-quantity
                :quant 2
                :unit (z6/year)))
   :time(z7/after
       :op1(z8/test-01
           :ARG1 z2
           :ARG2(z9/positive)
           :ARG3(z10/cocaine)
           :time(z11/game
                :name(z12/name
                    :op1 "Wimbledon")
                :time(z13/date-entity
                    :year 2007)))))
```

Error Type	Description	Example
Predicate Error	The predicate of the summary does not align with the information provided in the original document.	Hingis began a two-year ban after testing positive for cocaine at 2007 Wimbledon.
Entity Error	The primary arguments, or their attributes, associated with the predicate are incorrect.	Kournikova began a two-year ban after testing positive for cocaine at Wimbledon in 2007.
Circumstance Error	The supplementary details, such as location or time, that define the context of a predicate are incorrect.	Hingis began a two-year ban after testing positive for cocaine at Wimbledon in 2014.
Discourse Link Error	Errors arise from the improper connection of statements within the discourse, such as errors in temporal ordering or causal link.	Hingis began a two-year ban before testing positive for cocaine at Wimbledon in 2007.
Out of Article Error	The statement conveys information that is absent in the original document.	Hingis and Kournikova will team up again to play at this year's All England Club.

Figure 3: Typology of factual errors. Given the source document and reference summary, we apply five kinds of factual inconsistencies: predicate error, entity error, circumstance error, discourse link error, and out-of-article error. Each color represents the implementation of one kind of factual error from reference summary to perturbed summary.

2 Abstract Meaning Representations

Introduced by Banarescu et al. (2013), AMR is a representation language that effectively captures the essence of texts by encoding abstract-level semantic information, including named entities, negations, coreferences, and modalities. This intrinsic capability positions AMR as a critical asset for various semantic-related NLP tasks, such as summarization (Liao et al., 2018) and machine translation (Song et al., 2019). In our study, we harness the potential of AMR in the summarization factuality evaluation task by perturbing the graphs of factually consistent summaries. **Each perturbation reflects a factual error in summarization.** In AMR graphs, nodes symbolize entities and concepts, connected through different relational edges. Figure 1 provides an example of an AMR graph for a summary.

3 AMRFact

In AMRFact, we start by parsing reference summaries into AMR graphs and then introduce factual errors into these graphs. Subsequently, we back-translate the perturbed AMR graphs into text and use data filtering to acquire high-quality negative training samples. Finally, we combine the positive samples (reference summaries) with the negatives to train a classifier, which serves as our factuality evaluation metric. More specifically, the process of *generating* factually inconsistent summaries involves two primary states: introducing summary-level perturbations based on AMR and invalid negative data filtering. They are discussed in more detail in the following sections dedicated to explaining how factual errors are introduces (§3.1) and the criteria for choosing valid negative examples (§3.2).

3.1 AMR-based Summary Perturbations

Our approach targets five types of *factual inconsistencies*: predicate error, entity error, circumstance error, discourse link error, and out-of-article error, as categorized by Pagnoni et al. (2021)[1]. A detailed description of these factual errors is presented in Figure 3. Each type of errors can be generated by perturbing the AMR graph of a summary that is initially factually consistent, and then reconverting the perturbed graph back into a natural language summary. Specifically, we start with a source document D alongside its factually consistent summary S^+, which is then parsed into an acyclic AMR graph G^+. This graph is manipulated to form a new AMR graph G^- that embodies factual inconsistencies. For instance, as illustrated in Figure 2, we produce a predicate error by swapping the original predicate "work" to "leisure". The perturbed AMR graph G^- is then transformed into a factually inconsistent natural language summary S^- using an AMR-to-Text model (Ribeiro et al., 2021).

Predicate Error. Predicate errors occur when the predicate in a summary does not align with the information in the source document. We simulate this type of error based on two processes: (1) by

[1] We do not discuss or implement coreference errors because this work targets sentence-level summaries, and it is hard for them to contain pronouns or references with wrong or nonexisting antecedents.

adding or removing polarity and (2) through the substitution of a predicate with its antonym.

Entity Error. Entity errors manifest when the entities associated with a predicate in a summary are incorrectly attributed or erroneous. These errors are crafted through two principal sources: (1) by swapping the roles of the agent and the patient, which results in the misattribution of actions or characteristics, and (2) through the substitution of specific entities, such as names and numbers.

Circumstance Error. Circumstance errors in summaries emerge when there is incorrect or misleading information regarding the context of predicate interactions, specifically in terms of location, time, and modality. These errors are mainly created in two ways: (1) by intensifying the modality, which alters the degree of certainty or possibility expressed in the statement, and (2) by substituting specific entities like locations and times.

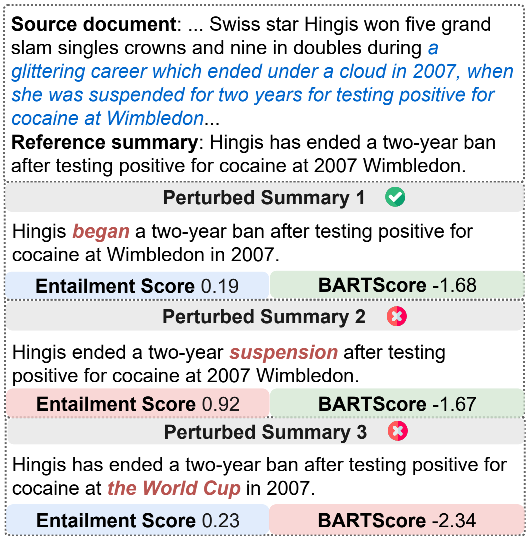
Discourse Link Error. Discourse link errors pertain to mistakes in the logical connections between various statements in a summary. We focus on two fundamental types of discourse links: (1) temporal order, which deals with the sequence of events, and (2) causality, which pertains to the cause-and-effect relationships between statements.

Out of Article Error. Summaries are expected to contain only information that can be inferred from the source document, and deviations from this rule need to be clearly identified. To create an "out of article" error, we follow a similar method as previously discussed, involving alterations in entities, times, or locations. However, in this instance, we intentionally introduce vocabulary not present in the original document.

3.2 Invalid Negative Data Filtering

In our data generation process, we noticed that certain generated negative examples either blatantly contradicted the source article's facts or failed to modify the semantics of the reference summary, inadvertently becoming *positive* examples These were deemed as *invalid* negative examples. We hypothesize that *training detection models on such invalid negative data could potentially impair their performance*. To address this issue, we introduce a module NEGFILTER specifically designed for filtering out invalid negative data.

A perturbed summary is considered *valid* only if it satisfy two essential criteria: (1) it must be sufficiently distinct from the original summary to avoid being mistaken for a mere variation, thereby ensuring it is not misclassified as a positive example; and (2) despite the introduced perturbations, it should maintain a discernible connection to the source document, without diverging excessively.



Source document: ... Swiss star Hingis won five grand slam singles crowns and nine in doubles during *a glittering career which ended under a cloud in 2007, when she was suspended for two years for testing positive for cocaine at Wimbledon*...
Reference summary: Hingis has ended a two-year ban after testing positive for cocaine at 2007 Wimbledon.

Perturbed Summary 1 ✓

Hingis ***began*** a two-year ban after testing positive for cocaine at Wimbledon in 2007.

Entailment Score 0.19 | **BARTScore** -1.68

Perturbed Summary 2 ✗

Hingis ended a two-year ***suspension*** after testing positive for cocaine at 2007 Wimbledon.

Entailment Score 0.92 | **BARTScore** -1.67

Perturbed Summary 3 ✗

Hingis has ended a two-year ban after testing positive for cocaine at ***the World Cup*** in 2007.

Entailment Score 0.23 | **BARTScore** -2.34

Figure 4: An example showing how our invalid negative data filtering module works. In the above three examples, only the first perturbed summary is valid since both of its entailment score and BARTSCORE satisfy the criteria described in §3.2.

To ensure the first criterion – *distinctiveness of generated summaries* – we utilize sentence-level NLI inspired by previous studies (Wan and Bansal, 2022; Huang et al., 2023a,b). Concretely, we employ a RoBERTa large model fine-tuned on the MNLI corpus (Liu et al., 2019; Williams et al., 2018)[2] to quantify the entailment score between an original summary (S^+) and its perturbed counterpart (S^-). Then, perturbed summaries that score exceed an empirically selected threshold (τ_1) in entailment scores are discarded, given their elevated likelihood of being inferred from S^+.

For the second criterion – *maintaining relevance to the source document* – we propose to use BARTSCORE (Yuan et al., 2021), fine-tuned on the CNN/DM dataset (Hermann et al., 2015), to assess the semantic alignment between the perturbed summary (S^-) and its corresponding source document (D). Similarly, perturbed summaries with a BARTSCORE falling below the empirically determined threshold (τ_2) are excluded due to their divergence from the source documents.

In summary, our proposed high-quality negative examples selection module NEGFILTER will take a source document (D), an original summary (S^+), and a perturbed summary (S^-). Only when both

[2] https://huggingface.co/roberta-large-mnli

criteria are satisfied, the generated factually inconsistent summaries (S^-) will be included into our negative training examples. Formally,

$$\mathcal{M}(D, S^+, S^-) = \textbf{True} \text{ if and only if}$$
$$\mathbb{1}[\mathcal{N}(S^+, S^-) < \tau_1] \cdot \mathbb{1}[\mathcal{B}(D, S^-) > \tau_2] = 1, \quad (1)$$

where N represents the entailment score, and B denotes BARTSCORE[3]. Figure 4 provides an illustration of our proposed selection module.

3.3 Detecting Factual Inconsistency

To learn to detect factual inconsistency, we fine-tune a `roberta-large` model pre-trained on the MNLI corpus on our augmented AMRFACT dataset, given that this model has been fine-tuned on a relevant task. The input to our model is a concatenation of the document premise D with a summary hypothesis (S^+ or S^-). Although our output linear layer performs a three-way classification, we focus only on the "entailment" and "contradiction" outputs. The training data paired with the reference summaries are labeled as "entailment", whereas those coupled with the perturbed summaries are labeled as "contradiction".

4 Experimental Settings

4.1 Datasets

Training Dataset We applied the AMRFACT data generation pipeline to the training split of the CNN/DM corpus (Hermann et al., 2015), compromising English news articles and their associated human-generated summaries. We created negative training data by breaking each reference summary into sentences. Each sentence is then perturbed using our AMR-based manipulation and processed by NEGFILTER. These selectively chosen perturbed summaries are combined with the positive examples (reference summaries) to form our training dataset. Eventually, our dataset consists of 13,834 training and 2,000 validation instances.

Evaluation Dataset Tang et al. (2023) introduced the benchmark **AGGREFACT**, which consolidates *nine* existing datasets focused on factuality for a finer-grained comparison of factuality assessment systems. They stratify the benchmark based on the underlying summarization model, categorized into FTSOTA, EXFORMER, and OLD according to their development timeline. FTSOTA encompasses state-of-the-art fine-tuned summarization models, such as BART (Lewis et al., 2020) and PEGASUS (Zhang et al., 2019). Models in the EXFORMER and OLD split were developed much earlier, such as Pointer-Generator (See et al., 2017). In line with Tang et al. (2023)'s recommendations, our performance reporting focuses on the FTSOTA split, as it most accurately reflects the challenges of unfaithfulness in the *most advanced* summarization models. Table 5 shows the statistics of AGGREFACT-FTSOTA.

4.2 Baseline Models

We compared AMRFACT with the following *non-LLM* methods. **FACTCC** (Kryscinski et al., 2020) is an entailment-based metric trained on synthetic data created through rule-based transformations of source document sentences. **DAE** (Goyal and Durrett, 2021) proposes an arc entailment approach that evaluates the factuality of each dependency arc of the generated summary independently with respect to the input article and then uses their aggregation as the overall score. Unlike previous work that reports the performance of DAE trained on human-annotated data, we opt for DAE-ENT, a variant of DAE trained on synthetic data, for fair comparison. **QUESTEVAL** (Scialom et al., 2021) introduces a QA-based metric that calculates factuality by aggregating answer overlap scores from queries derived from both the input article and the summary. **SUMMAC** (Laban et al., 2022) focuses on detecting factual inconsistencies by aggregating sentence-level entailment scores between the input document and summary sentences, with (**CONV**) or without (**ZS**) further fine-tuning. **QAFACTEVAL** (Fabbri et al., 2022) is a QA-based metric analogous to the precision-based component of QUESTEVAL and includes optimized question answering, generation, and answer-overlap components. **ALIGNSCORE** (Zha et al., 2023) learns a metric based on information alignment between two texts by integrating a diverse range of data sources. **FALSESUM** (Utama et al., 2022) is an entailment-based method that is trained on negative data produced by infilling masked spans of reference summaries with control code. For a fair comparison with entailment-based models, we retrain FACTCC and FALSESUM with `robera-large-mnli` as the base models.

In addition, we also compared with *LLM-based* methods. We tested **ChatGPT** (OpenAI, 2023) using different prompts (Luo et al., 2023; Wang et al., 2023): zero-shot binary rating (**ZS**), zero-shot binary rating with chain-of-thought (**CoT**), direct as-

[3] We set $\tau_1 = 0.9$ and $\tau_2 = -1.8$ for best performance.

	AggreFact-CNN-FtSota	AggreFact-XSum-FtSota	Avg
	Non-LLM-based		
DAE	65.0 ± 3.5	62.3 ± 1.9	63.7
QuestEval	70.2 ± 3.2	59.5 ± 2.7	64.9
SummaC-ZS	64.0 ± 3.8	56.4 ± 1.2	60.2
SummaC-Conv	61.0 ± 3.9	65.0 ± 2.2	63.0
QAFactEval	67.8 ± 4.1	63.9 ± 2.4	65.9
FactCC	57.6 ± 3.9	57.2 ± 1.7	57.4
FalseSum	50.5 ± 3.3	54.7 ± 1.9	52.6
AlignScore	67.0 ± 3.1	60.3 ± 1.9	63.7
	LLM-based		
ChatGPT-ZS	56.3 ± 2.9	62.7 ± 1.7	59.5
ChatGPT-COT	52.5 ± 3.3	55.9 ± 2.1	54.2
ChatGPT-DA	53.7 ± 3.5	54.9 ± 1.9	54.3
ChatGPT-Star	56.3 ± 3.1	57.8 ± 0.2	57.1
G-Eval	69.9 ± 3.5	**65.8 ± 1.9**	67.9
AMRFact (ours)	**72.3 ± 2.5**	64.1 ± 1.8	**68.2**

Table 1: Balanced binary accuracy (%) on the AggreFact-FtSota test set. We show 95% confidence intervals. Highest performance is highlighted in **bold**. The Avg score is computed by taking the arithmetic average of the performance on AggreFact-CNN-FtSota and AggreFact-XSum-FtSota.

sessment on a continuous scale from 0 to 100 (**DA**), and direct assessment on a discrete scale of one-to-five (**Star**). **G-Eval** (Liu et al., 2023) uses chain-of-thoughts and a form-filling template to produce a discrte scale of one-to-five. We implement it with GPT-4 Turbo (`gpt-4-1106-preview`).

4.3 Evaluation Criteria

We use balanced accuracy (Brodersen et al., 2010) as our evaluation metric. Then, we set a threshold for each model such that the threshold optimizes its performance on the validation set, following Laban et al. (2022). Therefore, our models will produce a binary label instead of a scalar value.

5 Results and Discussion

5.1 Main Results

The summarized results in Table 1 indicates that our model AMRFact sets a new benchmark on the AggreFact-FtSota test set. We report the performance on CNN/DM and XSum using separate thresholds, following Tang et al. (2023). AMRFact establishes a new state-of-the-art score of 72.3% on the CNN/DM split, outperforming the nearest competitor by a margin of 2.1%. On XSum split, our model's performance is competitive, tying with the highest score at 64.1%. With an overall average score of 68.2% across both datasets, we outperform the previous best model in the average by 2.3% over all non-LLM and LLM-based approaches. The improvements underscore the advantage of incorporating AMR in generating training data and the strategic selection of high-quality negative data through our data selection module.

	AggreFact-CNN-FtSota	AggreFact-XSum-FtSota	Avg
FactCC	57.6 ± 3.9	57.2 ± 1.7	57.4
+ Filtering	67.9 ± 2.3	63.8 ± 2.2	65.8
FalseSum	50.5 ± 3.3	54.7 ± 1.9	52.6
+ Filtering	52.3 ± 1.8	59.7 ± 2.2	56.0
AMRFact (ours)	**72.3 ± 2.5**	**64.1 ± 1.8**	**68.2**
- Filtering	64.4 ± 3.0	57.8 ± 2.0	61.1

Table 2: Balanced binary accuracy (%) on the AggreFact-CNN-FtSota and AggreFact-XSum-FtSota test set with or without our invalid negative data filtering module.

5.2 Impact of Our Filtering Module

To validate the *effectiveness* and *generalizability* of our invalid negative data filtering component NegFilter, we apply this module to the data generated by FactCC and FalseSum and re-train a RoBERTa on these newly filtered data. Additionally, we train another RoBERTa on our generated data without the filtering component. The results are summarized in Table 2. We observe a substantial performance gain when all three data generation methods incorporate the proposed invalid data filtering module. Specifically, FactCC's balanced accuracy improves from 57.6% to 67.9% on the CNN/DM and from 57.2% to 63.8% on XSum, while FalseSum sees an enhancement from 50.5% to 52.3% on CNN/DM and from 54.7% to 59.7% on XSum. These improvements affirm the effectiveness and generalizability of our approach in filtering invalid negative training data. Conversely, removing the filtering from our AMRFact data causes a decrease in balanced accuracy, dropping from 72.3% to 64.4% on CNN/DM and from 64.1% to 57.8% on XSum, resulting in an average drop from 68.2% to 61.1%. This stark contrast underlines the critical role of our invalid negative data filtering module in enhancing the performance of inconsistency detection models, supporting our hypothesis mentioned in §3.2 that **training inconsistency detection models on invalid negative data hurts the performance**.

5.3 Ablation Studies

In the next step, we focus on the impact of five specific perturbations on metric performance, using the AggreFact-FtSota dataset for an in-depth ablation analysis. This involves removing each perturbation from the AMRFact dataset to observe

	AggreFact-CNN-FtSota	AggreFact-XSum-FtSota	Avg
AMRFact	**72.3 ± 2.5**	64.1 ± 1.8	**68.2**
- Predicate Error	70.0 ± 3.1	62.0 ± 1.9	66.0
- Entity Error	70.8 ± 3.1	62.1 ± 2.1	66.5
- Circumstance Error	68.2 ± 2.9	61.6 ± 1.9	64.9
- Discourse Link Error	65.0 ± 2.7	61.7 ± 1.9	63.4
- Out of Article Error	66.8 ± 2.4	62.3 ± 2.0	64.6

Table 3: Balanced binary accuracy (%) on the AggreFact-Cnn-FtSota and AggreFact-XSum-FtSota test set *without* a specific error type.

how their absence influences the creation of negative examples. As outlined in Table 3, the results indicated a significant decrease in metric accuracy when any perturbation was excluded, highlighting their importance in enhancing metric precision.

The most notable finding is the substantial decline in performance upon removing discourse link error, suggesting that current models frequently struggle with this issue. By eliminating this perturbation, the model cannot access such negative examples during training, which significantly limits its ability to detect such inconsistencies during inference time. Conversely, omitting entity and out-of-article errors from the perturbations have a minimal impact on CNN/DM and XSum, respectively, indicating the relative robustness of current models against these specific issues.

5.4 Qualitative Analysis

The following qualitative analysis provides insights into our model's capability to produce coherent negative examples while ensuring extensive coverage of various error types.

Coherence To demonstrate that our approach produces more coherent negative summaries than entity-based baselines, we compare 200 summaries generated by AMRFact and FactCC. We use GPT-4 Turbo to assess the coherence of each summary on a scale of 1-5, where 1 indicates the least coherent and 5 means the most. The average coherence scores for AMRFact and FactCC are 3.01 and 2.24, respectively, highlighting the advantages of our approach in generating more coherent negative summaries compared to FactCC. We show the breakdown score in Figure 6.

Error-type coverage To verify that generation-based baselines suffer from the insufficient error-type coverage issue, we evaluate the error type distribution of the negative data produced by FalseSum by sampling its generated summaries and query GPT-4 Turbo to determine the error type within each summary. The findings revealed a notable gap: only 5 out of 1,000 summaries exhibited a Discourse Link Error. This scarcity of training data for this error type could lead to models trained on such data failing to detect these errors reliably. In Table 4, we show an example with a Discourse Link Error in which our model predicts the correct label and FalseSum fails, underscoring the issue of inadequate error-type coverage. The prompts used for both analyses are detailed in Appendix C.

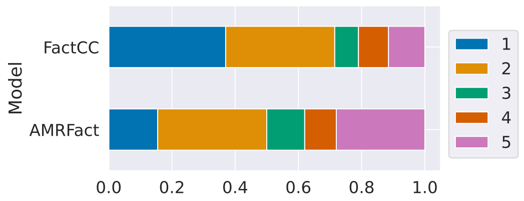


Figure 5: A breakdown of coherence scores for negative summaries produced by AMRFact and FactCC.

5.5 Remaining Challenges

To gain deeper insights into our model's limitations and remaining challenges, we conducted a detailed manual analysis. This involved examining a random sample of 50 examples from AggreFact-FtSota where our model incorrectly labeled factually inconsistent instances. The results of this analysis are meticulously detailed in Figure 6. From this thorough assessment, several key findings emerged. First, the most dominant type of error (30%) is the annotation error. This underscores the challenges in creating completely reliable benchmarks through crowd-sourced methods, a complexity also noted in the work of Laban et al. (2023). Second, errors stemming from out-of-article content were another significant source of inaccuracies. Intriguingly, upon further investigation, we discovered that a notable portion (8 out of 14) of these out-of-article errors were, in fact, factually accurate (Dong et al., 2022). For instance,

> Former Wales captain Martyn Williams says Dan Biggar's decision to sign a new four-year contract with Ospreys will benefit the region.

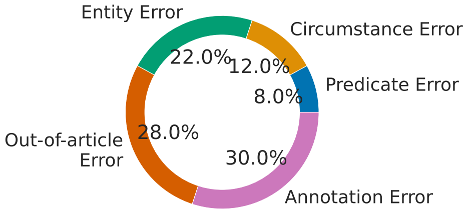


Figure 6: Distribution of the errors that our model fails to identify. The most dominant type of error is the annotation error. And errors stemming from out-of-article content were another significant source of inaccuracies.

Article: Lord Janner signed a letter saying he wanted to remain a peer just a week before he was ruled unfit to face child sex charges. Abuse campaigners last night angrily questioned why the suspected paedophile was able to remain in the House of Lords if he was too frail to be brought before court...

Summary: Lord Janner signed letter saying he wanted to remain a peer on April 9. Comes a week after he was ruled unfit to face child sex charges...

Table 4: An example with a discourse link error from AggreFact-FtSota where models trained on AMRFact successfully classify it as factual inconsistent with the input article but fails when trained on FalseSum.

The above summary was labeled as unfaithful because the corresponding source document (see Table 6) does not mention that Martyn Williams was a captain. Nonetheless, this information is factually correct, as evidenced by a Wikipedia search. This discrepancy suggests that our model, in this case, might have utilized its parametric knowledge learned during the pre-training phase of RoBERTa rather than the content of the article itself. To mitigate this issue, future research efforts can look into creating negative training examples that consist of out-of-article errors which, although truthful according to worldly facts, are unfaithful to the source document.

6 Related Work

6.1 Factual Consistency Metrics

Research on factual consistency metrics can be classified into QA-based and entailment-based approaches. QA-based metrics generally leverage question generation and answering components to evaluate factual consistency through the comparison of information units taken from summaries and their original sources (Wang et al., 2020; Scialom et al., 2021; Fabbri et al., 2022; Min et al., 2023). Entailment-based methods predict whether a summary is entailed by its source article with document-sentence entailment models trained on either synthetic (Kryscinski et al., 2020; Yin et al., 2021; Utama et al., 2022) or human-annotated data (Goyal and Durrett, 2021; Ribeiro et al., 2022; Chan et al., 2023). Besides, Gekhman et al. (2023) propose a method for generating synthetic data by annotating diverse model-generated summaries using an LLM. Alternatively, Laban et al. (2022) break documents and summaries into sentences and employ traditional sentence-level NLI models. Moreover, Feng et al. (2023) utilize facts extracted from external knowledge bases to improve the generalization across domains. Our approach employs an entailment-based method, capitalizing on AMR-based perturbations to optimize error-type coverage and synthetic data quality. We further improve data quality with NegFilter.

6.2 AMR-based Evaluation

Few studies have attempted to employ AMR for evaluation. Ribeiro et al. (2022) parse both the input documents and summaries into AMR graphs and train a classifier to identify erroneous edges in the summary graph. However, their approach relies on human-annotated data for effective training signals. On the contrary, we parse summaries into AMR graphs and create negative training data by applying perturbations to the graphs and converting the perturbed graphs back to summaries. Ghazarian et al. (2022) uses a similar approach to generate training data for a dialogue coherence evaluation metric. Notably, their perturbation techniques focused on coherence evaluation, and thus ill-suited for factual consistency evaluation. By contrast, we design our perturbation method by taking inspiration from common factual errors in summaries produced by state-of-the-art summarization systems.

7 Conclusion

We introduce a novel framework, AMRFact, which leverages Abstract Meaning Representations (AMRs) to generate factually inconsistent summaries for summarization factuality evaluation. Our method addresses the prevalent issues of low coherence and insufficient error-type coverage observed in prior entailment-based approaches. By parsing factually consistent reference summaries into AMR graphs and injecting controlled factual inconsistencies, we successfully create coherent but factually inconsistent summaries with broad error-type coverage. The introduction of the NegFilter module further enhances the quality of the generated negative samples. The effectiveness and generalizability of this approach are validated through its state-of-the-art performance on the AggreFact-FtSota dataset. Our contributions not only advance the field in generating high-quality, factually inconsistent summaries but also provide a scalable and efficient solution for enhancing the factuality evaluation of summarization systems. The results underscore the potential of AMR-based perturbations in improving the integrity and reliability of natural language generation.

8 Ethical Considerations

The metrics introduced in this study are based on models trained predominantly with English language documents. Consequently, they reflect the cultural nuances and perspectives common in English-speaking societies. It is important to acknowledge the potential presence of political and gender biases within these datasets. The models, and by extension, the metrics derived from them, might inadvertently perpetuate these biases. We have not rigorously evaluated these metrics for bias-related issues. Therefore, we advise users to be mindful of these limitations when applying these metrics, considering the potential for underlying biases in their use and interpretation. On a positive note, our methodology can serve as a valuable tool for detecting factually inconsistent summaries and making the summaries more factual.

9 Limitations

When employed effectively, the metrics outlined in this paper can serve as valuable tools for identifying errors in summarization models. However, it is crucial to recognize that these metrics are not infallible in detecting every factual inconsistency. This limitation should be considered when using these metrics to assess summaries for downstream applications. Moreover, our goal is to demonstrate the potential of using AMR-based perturbations for generating coherent yet intentionally factually inconsistent summaries. However, we acknowledge that the quality of our perturbations depends on pre-trained text-to-AMR parsers and AMR-to-Text generators. If these models are not strong, our summaries may suffer in quality, as discussed in §3.2. Therefore, it is essential to be aware of these constraints because factual inconsistencies in summaries can potentially propagate misinformation online. Therefore, while these metrics are helpful, their limitations in fully capturing factual inconsistencies must be acknowledged and managed carefully.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *LAW@ACL*.

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

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hou Pong Chan, Qi Zeng, and Heng Ji. 2023. Interpretable automatic fine-grained inconsistency detection in text summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6433–6444, Toronto, Canada. Association for Computational Linguistics.

Yue Dong, John Wieting, and Pat Verga. 2022. Faithful to the document or to the world? mitigating hallucinations via entity-linked knowledge in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1067–1082, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

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.

Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. 2023. Factkb: Generalizable factuality evaluation using language models enhanced with factual knowledge. *arXiv preprint arXiv:2305.08281*.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. Trueteacher: Learning factual consistency evaluation with large language models. *ArXiv*, abs/2305.11171.

Sarik Ghazarian, Nuan Wen, Aram Galstyan, and Nanyun Peng. 2022. DEAM: Dialogue coherence evaluation using AMR-based semantic manipulations. In *Proceedings of the 60th Annual Meeting of* *the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 771–785, Dublin, Ireland. Association for Computational Linguistics.

Ben Goodrich, Vinay Rao, Mohammad Saleh, and Peter J. Liu. 2019. Assessing the factual accuracy of generated text. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems*, December 7-12, 2015, Montreal, Quebec, Canada, pages 1693–1701.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online. Association for Computational Linguistics.

Kung-Hsiang Huang, Hou Pong Chan, and Heng Ji. 2023a. Zero-shot faithful factual error correction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5660–5676, Toronto, Canada. Association for Computational Linguistics.

Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2023b. Faking fake news for real fake news detection: Propaganda-loaded training data generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14571–14589, Toronto, Canada. 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, 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 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.

Fei-Tzin Lee, Christopher Kedzie, Nakul Verma, and Kathleen McKeown. 2021. An analysis of document graph construction methods for amr summarization. *ArXiv*, abs/2111.13993.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Kexin Liao, Logan Lebanoff, and Fei Liu. 2018. Abstract meaning representation for multi-document summarization. *arXiv preprint arXiv:1806.05655*.

Yang Liu, Dan Iter, Yichong Xu, Shuo Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *ArXiv*, abs/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.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *EMNLP*.

OpenAI. 2023. Chatgpt.

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.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Leonardo F. R. Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. FactGraph: Evaluating factuality in summarization with semantic graph representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3238–3253, Seattle, United States. Association for Computational Linguistics.

Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2021. Investigating pretrained language models for graph-to-text generation. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 211–227, Online. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. 2019. Semantic neural machine translation using AMR. *Transactions of the Association for Computational Linguistics*, 7:19–31.

Robyn Speer and Catherine Havasi. 2012. Representing general relational knowledge in ConceptNet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3679–3686, Istanbul, Turkey. European Language Resources Association (ELRA).

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. 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.

David Wan and Mohit Bansal. 2022. FactPEGASUS: Factuality-aware pre-training and fine-tuning for abstractive summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1010–1028, 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.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *ArXiv*, abs/2303.04048.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. In *Annual Meeting of the Association for Computational Linguistics*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *CoRR*, abs/1912.08777.

A Error Typology Detailed Description

Predicate Error. Predicate errors occur when the predicate in a summary does not align with the information in the source document. We simulate this type of error based on two processes: (1) by adding or removing polarity and (2) through the substitution of a predicate with its antonym. By directly adding or removing this argument, we change the negation in the sentence. Another approach is the antonym substitution. Here, we replace the concepts with their antonyms that holding *Antonym*, *NotDesires*, *NotCapableOf*, and *NotHasProperty* relations in ConceptNet (Speer and Havasi, 2012), and therefore modify the sentence-level relations. The strengths of the AMR-to-Text model become evident when adaptations, like incorporating *to*, are coupled with verb substitutions to enhance summary coherence.

Entity Error. Entity errors manifest when the entities associated with a predicate in a summary are incorrectly attributed or erroneous. These errors are crafted through two principal sources: (1) by swapping the roles of the agent and the patient, which results in the misattribution of actions or characteristics, and (2) through the substitution of specific entities, such as names and numbers. In AMR graphs, the clear demarcation between agent (*ARG0*) and patient (*ARG1*) allows for straightforward swaps. For named entity modifications, we employ a rule-based approach using the SpaCy NER tagger (Honnibal et al., 2020) to extract the entities holding *PERSON*, *ORG*, *NORP*, *FAC*, *GPE*, *PRODUCT*, *WORK_OF_ART*, *EVENT*, *PERCENT*, *MONEY*, *QUANTITY*, and *CARDINAL* labels and then randomly select an entity from the summary to be replaced by a different entity with the same label from the same source document. After replacements, the AMR-to-text model effectively adjusts the summary segments, ensuring outputs that are both grammatically correct and naturally phrased. For instance, after swapping agent and patient, the nested subgraphs related to the arguments will be adjusted accordingly.

Circumstance Error. Circumstance errors in summaries emerge when there is incorrect or misleading information regarding the context of predicate interactions, specifically in terms of location, time, and modality. These errors are mainly created in two ways: (1) by intensifying the modality, which alters the degree of certainty or possibility expressed in the statement, and (2) by substituting specific entities like locations and times. We strengthen the modality by replacing the concepts that controls modals (*i.e.*, *permit-01*, *possible-01*, *likely-01*, *recommend-01*, *wish-01* → *obligate-01*). We emphasize modality strengthening over its inverse to avoid generating truth-conditionally compatible sentences. For entity substitution, we adopt a rule-based strategy with the SpaCy NER tagger (Honnibal et al., 2020) to extract the entities holding *LOC*, *DATE*, and *TIME* labels and then randomly select an entity from the summary to be replaced by a different entity with the same label from the same source document. The advantages of the AMR-to-Text model become apparent such as it introduces verb adaptations, enhancing the fluency of the summary due to the modality shift.

Discourse Link Error. Discourse link errors pertain to mistakes in the logical connections between various statements in a summary. We focus on two fundamental types of discourse links: (1) temporal order, which deals with the sequence of events, and (2) causality, which pertains to the cause-and-effect relationships between statements. To manipulate temporal ordering, we change *before* to *after*, *after* to *before*, and *now* to either *before* or *after* by recognizing their presence in a *time* argument. For causality modifications, we alter argument structures associated with *cause-01*, either at the root or as a modifier, effectively reversing the causal relationship. This includes transitions such as interchanging *because* with *therefore*. Utilizing the AMR-to-Text model ensures that the generated summary remains coherent and grammatically correct based on the perturbations.

Out of Article Error. Summaries are expected to contain only information that can be inferred from the source document, and deviations from this rule need to be clearly identified. To create an "out of article" error, we follow a similar method as previously discussed, involving alterations in entities, times, or locations. However, in this instance, we intentionally introduce vocabulary not present in the original document. Moreover, we propose the integration of irrelevant sources into AMR graphs by selecting AMR elements like concepts, *ops*, and *ARGs*, and substituting them with items from unrelated documents, coherence is deliberately disrupted. The strength of the AMR-to-Text model is showcased in this context, as it introduces adaptive

		Polytope	SummEval	FRANK	Wang'20	CLIFF	Goyal'21	Cao'22	Total
Cnn	val	34	200	75	-	150	-	-	459
	test	34	200	175	-	150	-	-	559
Xsum	val	-	-	-	120	150	50	457	777
	test	-	-	-	119	150	50	239	558

Table 5: Statistics of AggreFact-Cnn-FtSota and AggreFact-Xsum-FtSota. These two subsets consist of Polytope (Huang et al., 2020), SummEval (Fabbri et al., 2021), FRANK (Pagnoni et al., 2021), Wang'20 (Wang et al., 2020), CLIFF (Cao and Wang, 2021), Goyal'21 (Goyal and Durrett, 2021), Cao'22 (Cao et al., 2022).

The 26-year-old fly-half has agreed a new deal which could keep him in Wales until the 2019 World Cup. Williams, who played 100 times for Wales, believes it will help Ospreys keep and recruit players. "It's fabulous news for Welsh rugby and the Ospreys in particular," he said on BBC Wales' Scrum V TV programme. "Not only is Dan committing himself to Wales for the next four years, but it helps the Ospreys with recruitment for the next couple of seasons. "If players were looking to sign and come to the Ospreys if they can see Dan Biggar is going to be there for the next three or four seasons that helps them as well." Biggar's current deal was due to expire at the end of this season. He is the first of Wales' 17 dually contracted players to re-sign on the deals which are 60% funded by the WRU and 40% by the region. In addition to potentially attracting new players to the region, Biggar's decision to stay may help negotiations with his Ospreys and Wales colleagues scrum-half Rhys Webb and second row Alun Wyn Jones. Both players' contracts expire in the summer of 2016, with Ospreys skipper Jones saying in November that he was still weighing up his options. Scarlets are in talks with Wales centre Scott Williams over extending his dual contract, and have secured the return of British and Irish Lions centre Jonathan Davies from Clermont Auvergne next season.

Table 6: The source document corresponding to the factual but unfaithful summary mentioned in §5.5.

elements, like *with*, in tandem with new verb replacements, ensuring the summary retains fluency.

Generation. Given the source document D alongside its factually consistent summary S^{+}, the generation module first applies text-to-AMR models to the summary and translates it into directed and acyclic AMR graph G^{+}. Next, our contronlable generation model $\mathcal{G}$ injects factual errors specified above into AMR graph G^{+}, and outputs manipulated AMR graph G^{-} containing factually inconsistent information. We then back-translate the manipulated AMR graphs G^{-} into summary text S^{-} to be served as negative examples for training the text-based factuality evaluator.

B Data Statistics

Table 5 shows the statistics of AggreFact-FtSota.

C Prompt Details

We show the prompt used to analyze the error distribution for FalseSum in Table 7 and coherence evaluation in Table 8.

You will be given one reference summary and one generated summary written for a news article.
Your task is to determine the type of factual error in the generated summary with regard to the article.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Types of Errors:

1. Predicate Error: The predicate of the summary does not align with the information provided in the original document.
2. Entity Error: The primary arguments, or their attributes, associated with the predicate are incorrect.
3. Circumstance Error: The supplementary details, such as location or time, that define the context of a predicate are incorrect.
4. Discourse Link Error: Errors arise from the improper connection of statements within the discourse, such as errors in temporal ordering or causal link.
5. Out of Article Error: The statement conveys information that is absent in the original document.

Evaluation Steps:

1. Read the news article carefully and identify the main topic and key points.
2. Read the generated summary and compare it to the news article.
3. Determine the type of errors based on our definition.

Example:

Source Text:
[Document]

Reference Summary:
[Reference Summary]

Generated Summary:
[Generated Summary]

Evaluation Form (error type ONLY):

- Error Type:

Table 7: The prompt to GPT-4 Turbo for determining the error type of a non-factual summary.

You will be given one summary written for a news article.
Your task is to rate the summary on one metric.
Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:
Coherence (1-5). Here, we focus on "within sentence coherence". It involves ensuring that the components of a single sentence – such as subjects, verbs, objects, and other elements – are logically and grammatically connected, making the sentence clear and understandable.
Evaluation Steps:

1. Read the summary carefully.
2. Evaluate the summary based on the evaluation criteria.
3. Assign a score for coherence on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.

Generated Summary:
[Generated Summary]

Evaluation Form (scores ONLY):
- Coherence:

Table 8: The prompt to GPT-4 Turbo for determining the coherence of a generated summary.