FIZZ: Factual Inconsistency Detection by Zoom-in Summary and Zoom-out Document

Joonho Yang[1], **Seunghyun Yoon**[2], **Byeongjeong Kim**[1], **Hwanhee Lee**[1†]
[1]Department of Artificial Intelligence, Chung-Ang University, [2]Adobe Research, USA
{plm3332, michael97k, hwanheelee}@cau.ac.kr, syoon@adobe.com

Abstract

Through the advent of pre-trained language models, there have been notable advancements in abstractive summarization systems. Simultaneously, a considerable number of novel methods for evaluating factual consistency in abstractive summarization systems has been developed. But these evaluation approaches incorporate substantial limitations, especially on refinement and interpretability. In this work, we propose highly effective and interpretable factual inconsistency detection method **FIZZ** (**F**actual **I**nconsistency Detection by **Z**oom-in Summary and **Z**oom-out Document) for abstractive summarization systems that is based on fine-grained *atomic facts* decomposition. Moreover, we align *atomic facts* decomposed from the summary with the source document through adaptive granularity expansion. These *atomic facts* represent a more fine-grained unit of information, facilitating detailed understanding and interpretability of the summary's factual inconsistency. Experimental results demonstrate that our proposed factual consistency checking system significantly outperforms existing systems. We release the code at https://github.com/plm3332/FIZZ.

1 Introduction

With the development of pre-trained language models, abstractive summarization systems using these language models have made remarkable progress in generating fluent and natural summarizations (Chang et al., 2023). However, one of the notable challenges these systems confront is the hallucination, causing language models to generate summaries that are factually inconsistent with the given article (Maynez et al., 2020; Kryscinski et al., 2020; Tam et al., 2023; Zhang et al., 2023). Recognizing the significance of this issue, various evaluation metrics have been introduced to detect these errors, starting from traditional methods like ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020) to a large number of advanced metrics (Goyal and Durrett, 2020, 2021; Scialom et al., 2021; Fabbri et al., 2022; Laban et al., 2022; Luo et al., 2023; Zha et al., 2023; Wang et al., 2023a). Especially, many of the recent works (Laban et al., 2022; Schuster et al., 2022; Zha et al., 2023) adopted sentence level evaluation using Natural Language Inference (NLI) systems for factual consistency checking.

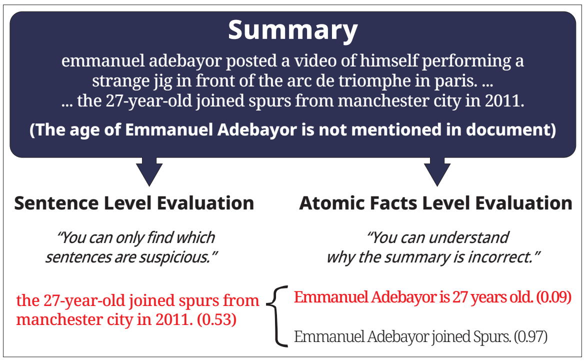

Summary
emmanuel adebayor posted a video of himself performing a strange jig in front of the arc de triomphe in paris. ...
... the 27-year-old joined spurs from manchester city in 2011.
(The age of Emmanuel Adebayor is not mentioned in document)
Sentence Level Evaluation
"You can only find which sentences are suspicious."
Atomic Facts Level Evaluation
"You can understand why the summary is incorrect."
the 27-year-old joined spurs from manchester city in 2011. (0.53)
Emmanuel Adebayor is 27 years old. (0.09)
Emmanuel Adebayor joined Spurs. (0.97)

Figure 1: Comparison between sentence level evaluation and atomic facts level evaluation. The numbers in parentheses represent the maximum NLI entailment scores obtained by comparing each sentence and atomic fact with the source document on a sentence-wise basis.

Although these studies have shown a certain level of performance in summary evaluation, they still exhibit significant deficiencies in accuracy. Additionally, they substantially lack in interpretability, an area crucial for further development in the field of summarization factual consistency detection. As shown in Figure 1, sentence level evaluation often fails to check the details of the various facts in each sentence, resulting in lower accuracy and lower interpretability. Furthermore, we find that pair-wise single sentence level evaluation is vulnerable to summary evaluation that requires multi-sentence reasoning. In addition, expressions such as pronouns in sentences can lead the NLI system to

[†] Corresponding author.

make incorrect judgments in single sentence level evaluation.

In this paper, we propose an interpretable summarization factual inconsistency detection system, FIZZ, which overcomes the issues of previous sentence level NLI-based evaluation. As in Figure 2, FIZZ first resolves coreferences in both the source document and the generated summary. Subsequently, we decompose this coreference resolved summary into *atomic facts*, which is an approach that zooms in the summary. This *atomic fact* can be considered a more fine-grained information unit embedded within the text than a sentence at a broad level. As in the *atomic fact* examples in Figure 1, a single sentence from the summary can be segmented into two or more distinct units of information. This approach allows for a more detailed analysis of textual information, which is crucial for evaluating the factuality of generated text. Using these *atomic facts*, we check the consistency of each *atomic fact* against the source document using an NLI model. As highlighted in Figure 1, factual inconsistencies that cannot be detected at the sentence level can be identified through evaluation at this atomic fact level with higher interpretability. Also, we propose a granularity expansion method that can adaptively increase the number of context sentences when verifying the consistency of each atomic fact. Through this way of zooming out the document, we efficiently check the consistency of certain atomic facts that require multi-sentence level reasoning.

Experimental results show that our proposed system FIZZ achieves state-of-the-art performance on the AGGREFACT (Tang et al., 2023) benchmark dataset. FIZZ exhibits high interpretability by utilizing *atomic facts*. Furthermore, We have tested on various LLMs to implement atomic fact generation task and identified the best model suited for this task. Additionally, our analysis shows that flexibly increasing the granularity choice of the source document significantly enhances accuracy.

2 Related Work

Summarization Factual Consistency Evaluation A multitude of metrics designed to evaluate summarization factual consistency are currently being refined by leveraging NLP pipelines originally developed for disparate tasks, including QA-based evaluation, parsing-based methods, LLM-based prompting, and NLI-based metrics.

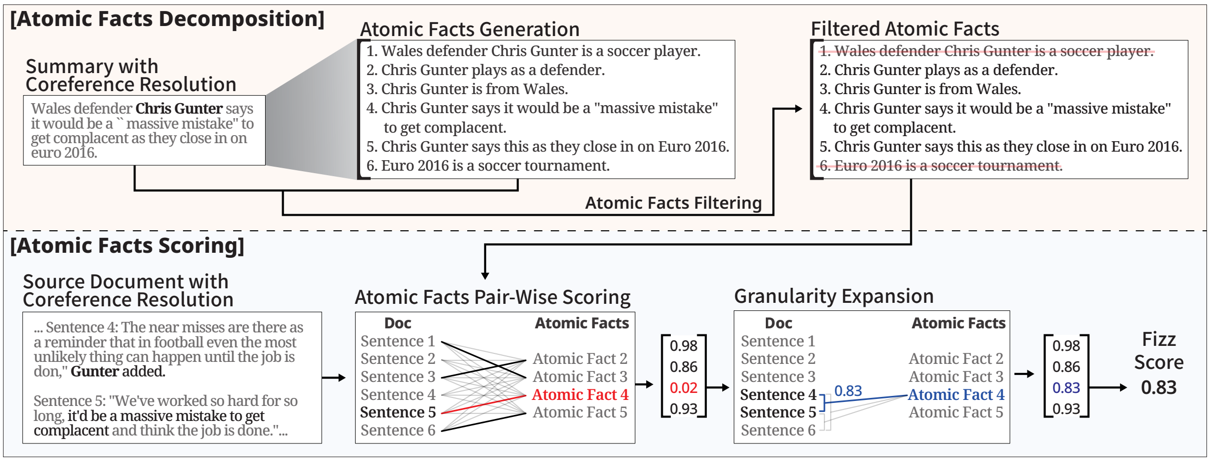
QA-based methods involve two steps of question generation (QG) and question answering(QA). While FEQA (Durmus et al., 2020) generate questions with the summary as the source, QUESTEVAL (Scialom et al., 2021) and QAFACTEVAL (Fabbri et al., 2022) generate questions with both the summary and the document.

Parsing-based methods discover relationships by employing syntactic parsing process, thereafter calculating the proportion of summary-derived relations that align with those extracted from source documents. Goodrich et al. (2019) extract relation tuples for the evaluation. DAE (Goyal and Durrett, 2020, 2021) propose utilizing a dependency arc between the entities and the relationship.

There is a growing trend for using LLMs like ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023) on summarization factual consistency checking (Luo et al., 2023; Chen et al., 2023; Wang et al., 2023a; Gekhman et al., 2023; Yang et al., 2024). Initially, Luo et al. (2023) explores ChatGPT's ability in evaluating factual consistency for text summarization with zero-shot prompting. Yang et al. (2024) extend the work by excluding irrelevant sentences from both documents before providing prompts to GPT-4.

SUMMAC (Laban et al., 2022) re-visit NLI-based models and granularity choice for inconsistency detection in summarization. ALIGNSCORE (Zha et al., 2023) develops an alignment system, incorporating a summarization consistency checking metric and an NLI model, which has been trained across a diverse array of tasks that can be *aligned* with NLI. The recently proposed method, FENICE (Scirè et al., 2024), also aligns decomposed *atomic facts* with several document sentences, but it lacks interpretability on summary side. Our proposed system, FIZZ, is also based on NLI. However, unlike the aforementioned systems, which mostly compare the summary at the sentence level, FIZZ conducts comparisons at a more fine-grained atomic fact level with high interpretability.

Atomic Facts Generation To the best of our knowledge, van Halteren and Teufel (2003) pioneered the introduction of an atomic information unit, named a *factoid*, within the field of summarization evaluation. Building on this foundational work, Nenkova and Passonneau (2004) proposed the Pyramid method, a manual evaluation protocol for summarization that employs *Summarization Content Units* (SCUs), also referred to as *Se-*



Wales defender **Chris Gunter** says it would be a `` massive mistake" to get complacent as they close in on euro 2016.

1. Wales defender Chris Gunter is a soccer player.
2. Chris Gunter plays as a defender.
3. Chris Gunter is from Wales.
4. Chris Gunter says it would be a "massive mistake" to get complacent.
5. Chris Gunter says this as they close in on Euro 2016.
6. Euro 2016 is a soccer tournament.

~~1. Wales defender Chris Gunter is a soccer player.~~
2. Chris Gunter plays as a defender.
3. Chris Gunter is from Wales.
4. Chris Gunter says it would be a "massive mistake" to get complacent.
5. Chris Gunter says this as they close in on Euro 2016.
~~6. Euro 2016 is a soccer tournament.~~

... Sentence 4: The near misses are there as a reminder that in football even the most unlikely thing can happen until the job is don," **Gunter** added.

Sentence 5: "We've worked so hard for so long, **it'd be a massive mistake to get complacent** and think the job is done."...

Figure 2: Overall flow of FIZZ. The pipeline begins by applying coreference resolution to both the summary and the document. Atomic facts are then decomposed from the summary using an LLM. These atomic facts are filtered and subsequently scored against the document. The scores are refined through granularity expansion. The ultimate score is defined by choosing the minimum score.

mantic Content Units. This innovative approach has inspired a significant body of subsequent research (Harnly et al., 2005; Shapira et al., 2019; Gao et al., 2019; Bhandari et al., 2020; Zhang and Bansal, 2021). Liu et al. (2023) referred to these elementary information units as *Atomic Content Unit*, or *Atomic Facts*. However, the realm of these investigations is primarily concentrated on assessing summarization itself via the examination of *atomic facts* crafted by human annotators[1].

In the scope of hallucination detection and fact verification for text generated by models, there has been a recent initiative to employ LLMs to create *atomic facts*. FACTSCORE (Min et al., 2023) utilize InstructGPT (Ouyang et al., 2022) for the creation of *atomic facts*. Following this work, FACTOOL (Chern et al., 2023) introduce a fact verification pipeline that leverages fine-grained information units generated by ChatGPT, referred to as *claims*. In this study, we present a novel method FIZZ leveraging atomic semantic unit, from now on called *atomic fact*, in the domain of summarization factual inconsistency detection.

3 FIZZ

The overall flow of our proposed system FIZZ is presented in Figure 2. Our method first begins with the application of a coreference resolution model to a given $(document, summary)$ pair, resulting in a new pair of texts $(document, summary)$ where coreferences have been resolved (Section 3.1). Following this, we proceed to generate *atomic facts* from the coreference-resolved summary leveraging LLMs as a zooming-in approach for the summary (Section 3.2). Using the generated atomic facts, we compute the score of each *atomic fact* with the NLI system (Section 3.3). Finally, we propose a granularity expansion method, which is a way of zooming out the documents, to compute the score for the summaries that contain high abstractiveness more accurately.

3.1 Coreference Resolution

To enhance the entailment recognition capabilities of NLI models, FIZZ first conducts centered around coreference resolution in both document and summary texts. The motivation behind this approach is driven by the inherent limitations observed in NLI models when processing texts with pronouns. Specifically, we find that NLI models tend to struggle with recognizing entailment when presented with *premises* and *hypotheses* that contain the same content but differ in their use of pronouns and explicit entity names. To address this challenge, FIZZ employs pronoun resolution in summaries by analyzing them on a sentence-by-sentence basis to extract atomic facts. This strategy not only facilitates a more granular understanding of the summary content but also avoids the limited context length in LLMs.

Furthermore, applying pronoun resolution to the document text ensures that the entities are explicitly named, aligning the *premise* more closely with the *hypothesis*. By resolving coreferences in both

[1] We note that Zhang and Bansal (2021) generated SCUs with semantic role labeling.

documents and summaries, our approach aims to bridge the gap between pronoun use and explicit entity naming, thereby improving the performance of NLI models in entailment tasks. This dual focus on both document and summary texts underscores the comprehensive nature of our strategy to bolster the accuracy and reliability of NLI models in handling a variety of linguistic expressions.

Formally, given a document D and its summary S, we define coreference resolution as f_{coref}, which makes:

$$D' = f_{\text{coref}}(D), \quad S' = f_{\text{coref}}(S) \qquad (1)$$

where D' and S' are coreference resolved texts of D and S, respectively.

3.2 Atomic Facts Decomposition

Atomic Facts Generation As demonstrated in Figure 1, sentence level evaluation of summaries can often yield inaccurate results. Therefore, we propose a method that evaluates the factuality of summaries at a more fine-grained level, specifically at the level of *atomic facts* as exemplified in Figure 2. By employing *atomic facts*, which are highly detailed units of information, FIZZ considerably enhances interpretability.

The definition of an *atomic fact* differs across studies, primarily due to the inherently subjective nature of this concept. We propose our own definition of an *atomic fact* that is designed to align with and complement the nature of NLI models. Building upon Bhandari et al. (2020), we specify further that an *atomic fact* is short and concise, containing no more than two or three entities, with person entities specifically resolved any of coreferences.

We generate atomic facts from summaries at the sentence level after resolving coreferences. This strategy for atomic fact generation not only increases the quantity of atomic facts but also substantially augments the generated summary's pool of information. To extract atomic facts from the summaries, we input prompts into the LLM that include both a task description and a sentence-level summary, as exemplified in Table 10. This approach systematically decomposes each sentence in the summary into individual atomic facts, facilitating a comprehensive extraction and representation of information. The coreference resolved summary $S' = \{s'_j\}_{j=1}^{N}$, where s'_j represents the j^{th} sentence in S' and N the total number of sentences in S', could be decomposed to a set of atomic facts $A' = \{a'_k\}_{k=1}^{L}$, with L denotes the total number of sentences in A'.

Algorithm 1 Filtering Out Incorrect Atomic Facts

```
Input: An NLI model $\mathcal{M}$; coreference resolved summary $S' = \{s'_j\}_{j=1}^{N}$; decomposed atomic facts $A' = \{a'_k\}_{k=1}^{L}$.
Initialize: set $A_{filtered} = \phi$
for $k = 1, 2, \ldots, L$ do
    for $j = 1, 2, \ldots, N$ do
        $(e_{j,k}, c_{j,k}, n_{j,k}) \leftarrow \mathcal{M}(s'_j, a'_k)$
        if $\max(e_{j,k}, c_{j,k}, n_{j,k})$ is $e_{j,k}$ then
            Append $a'_k$ to $A_{filtered}$.
        end if
    end for
end for
Output: A set of atomic facts $A_{filtered}$.
```

Atomic Facts Filtering One significant issue with atomic facts generated by LLMs is that these facts are often produced not from the content of summaries themselves but from the pretrained knowledge embedded within the LLMs. For example, when we decompose the sentence of the summary *"The mass, which has risen some 50ft above sea level, measures roughly 1,000 - 1,640ft long, and 100ft wide"*, the decomposed atomic facts contain an atomic fact *"The mass is a noun"*. Such atomic facts may not align with either the summaries or the documents and can significantly influence the scoring method described in Section 3.3. Consequently, the exclusion of these atomic facts becomes a necessary step in our process.

Hence, we utilize an NLI model to filter out incorrect atomic facts. Our approach leverages the probabilistic distribution of the NLI model, which categorizes outcomes into three types: *Entailment* (E), *Contradiction* (C), and *Neutral* (N). In the filtering process, we set the summary S' as the *premise*, and the atomic fact A' as the *hypothesis*. We filter out atomic facts that exhibit exceptionally low *entailment* scores. We outline the detailed procedure of the atomic facts filtering process in Algorithm 1.

3.3 Atomic Facts Scoring

Atomic Facts Pair-Wise Scoring To compute the score for each atomic fact of the summaries, FIZZ first decomposes the coreference resolved document into sentences. We split the document D' into M sentences and the filtered atomic facts $A_{filtered}$ into N sentences, formulating $D' = \{d'_i\}_{i=1}^{M}$ and $A_{filtered} = \{a_k\}_{k=1}^{L}$, respectively. We use each (d_i, a_k) as an input for an NLI model, positioning the generated atomic fact as the *hy-*

pothesis and the sentence of the document as the *premise*.

Finally, we assign scores to each atomic fact based on the `maximum` *entailment* score obtained through comparison with every sentence in the document. The atomic fact *entailment* scores $E = \{e_{i,k}\}$, where $1 \leq i \leq M$ and $1 \leq k \leq L$, are computed to a vector $\mathbf{T}$:

$$\begin{gathered} \mathbf{t}_k = \max_{1 \leq i \leq M} e_{i,k} \\ \mathbf{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_L\} \end{gathered} \tag{2}$$

Adaptive Granularity Expansion Summaries generated by *abstractive* summarization systems contain a high degree of *abstractiveness*. This *abstractiveness* occurs when content spread across multiple sentences in the document is condensed into one or two sentences in the summary. To accurately detect factual inconsistencies within such summaries, it is necessary to zoom out and examine multiple sentences across the source document. Furthermore, several studies have demonstrated that considering multiple sentences from the document leads to better accuracy (Laban et al., 2022; Glover et al., 2022).

We aim to identify scores where $\max(e_k, c_k, n_k)$ is *not* e_k from the $\mathbf{T}$. For atomic facts associated with these scores, we further increase the granularity of the document and perform computation once again. We incrementally increase the granularity starting from the document sentence d_i that contributed to each identified score, limiting the granularity at a maximum of three sentences (d_{i-1} + d_i, d_i + d_{i+1}, d_{i-2} + d_{i-1} + d_i, d_i + d_{i+1} + d_{i+2}, d_{i-1} + d_i + d_{i+1}). Subsequently, we re-calculate the scores within this expanded context and replace the original scores with the `maximum` value observed among the re-calculated scores and the original. As a result, the vector $\mathbf{T}$ is transformed into $\mathbf{T}^*$ as certain scores are replaced by new scores. Detailed information on this procedure is provided in Algorithm 2.

The final score is then determined by the `minimum` score within vector $\mathbf{T}^*$, enabling a highly interpretable evaluation:

$$\textit{FIZZ score} = \min(\mathbf{T}^*) \tag{3}$$

4 Experiments

4.1 Experimental Setups

In our experiments, we leverage MT5 (Bohnet et al., 2023) for coreference resolution, which returns with the identification of clusters referring to the same entities. With these clusters, we further implement rule-based pronoun substitution strategies to generate coreference resolved texts. For atomic fact decomposition, the Orca-2 model (Mitra et al., 2023) is utilized. Additionally, this work adopts the same off-the-shelf NLI model as implemented in SUMMAC (See Appendix D for more details).

Algorithm 2 Scoring with Document Granularity Expansion

Input: An NLI model $\mathcal{M}$; coreference resolved document $D' = \{d'_i\}_{i=1}^{M}$; decomposed atomic facts $A' = \{a'_k\}_{k=1}^{L}$.
Initialize: $\mathbf{T}^* = \phi$; Max granularity size $gran = 3$.

```
Define $\mathcal{C}(D, g)$ = list of subsets of $D$ with size of $g$.
Define $\mathcal{F}(\mathcal{C}(D, g))$ which returns whether $\mathcal{C}(D, g)$ is a consecutive list.
Define $\mathcal{D}(\mathcal{C}(D, g))$ = list of document sentences in index list in $\mathcal{C}(D, g)$.
for $k = 1, 2, \ldots, L$ do
    set $E = \phi$
    for $i = 1, 2, \ldots, M$ do
        $(e_{i,k}, c_{i,k}, n_{i,k}) \leftarrow \mathcal{M}(d'_i, a'_k)$
        Append $e_{i,k}$ to $E$.
    end for
    $m_{idx} = E.index(\max(E))$
    if $\max(e_{i,k}, c_{i,k}, n_{i,k})$ is not $e_{i,k}$ then
        set $D_{idx} = [0, \ldots, M-1]$
        set $D_{expanded} = \phi$
        for $g = 1, 2, \ldots, gran + 1$ do
            if $m_{idx}$ in $\mathcal{C}(D_{idx}, g)$ and $\mathcal{F}(\mathcal{C}(D_{idx}, g))$ then
                Extend $\mathcal{C}(D_{idx}, g)$ to $D_{expanded}$.
            end if
        end for
        set $E_{expanded} = \phi$
        for $d_{expanded} \in \mathcal{D}(D_{expanded})$ do
            $(e, c, n) \leftarrow \mathcal{M}(d_{expanded}, a'_k)$
            Append $e$ to $E_{expanded}$.
        end for
        Append $\max(E_{expanded})$ to $\mathbf{T}^*$.
    else
        Append $e_{i,k}$ to $\mathbf{T}^*$.
    end if
end for
```

Output: vector $\mathbf{T}^*$ with maximum entailment scores from each atomic fact.

4.2 Benchmark Datasets

We use AGGREFACT (Tang et al., 2023) benchmark dataset, a comprehensive aggregation of 9 leading summary factual consistency detection datasets currently available. AGGREFACT is stratified into three distinct splits, namely FTSOTA, EXFORMER, and OLD, with each split containing its own validation and test sets. We standardize the evaluation as a binary classification and choose the best threshold from the validation set following SummaC. Finally, we apply this threshold to the test set and report the balanced accuracy score, considering the imbal-

	AGGREFACT-CNN-FTSOTA	AGGREFACT-XSUM-FTSOTA	AVG
DAE	65.4±4.4	**70.2**±2.3	67.8
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
AlignScore	62.5±3.3	69.6±1.7	66.1
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
FactScore	60.8±3.2	68.0±2.0	64.4
FacTool	49.3±3.5	59.0±2.0	54.2
FIZZ (Ours)	**72.6**±3.0	69.3±1.9	**71.0**
$w/o\ GE$	72.2±2.8	66.3±1.9	69.3
$w/o\ Filtering$	64.7±3.3	70.0±1.8	67.4
$w/o\ AF$	63.6±2.9	65.8±2.0	64.7

Table 1: Balanced accuracy using a single threshold with 95% confidence intervals on the AGGREFACT-FTSOTA split dataset. Highest performance is highlited in **bold**, and the second highest is underlined.

ance in the dataset.

4.3 Baselines

We adopt all of the baselines of AGGREFACT dataset: DAE (Goyal and Durrett, 2020, 2021), QuestEval (Scialom et al., 2021), SummaC-ZS and SummaC-Conv (Laban et al., 2022), QAFactEval (Fabbri et al., 2022), ChatGPT-ZS and ChatGPT-CoT (Luo et al., 2023), ChatGPT-DA and ChatGPT-Star (Wang et al., 2023a). Also, we report the results with AlignScore (Zha et al., 2023), which is a recently introduced system for checking the factual consistency of summaries based on NLI. Additionally, we incorporate FACTSCORE (Min et al., 2023) and FACTOOL (Chern et al., 2023) in our baselines. These methods decompose generated texts into *atomic facts* and then retrieve corresponding entries from a given knowledge base, such as Wikipedia, to evaluate the factuality of the generated context. For the purpose of verification, we assume the availability of this knowledge base, which we use as the source document to assess summary factual consistency. In FACTSCORE, we employ a **No-context LM** for factual verification. This approach operates on a QA basis, assessing whether *atomic facts* are true or false with respect to the source document. In FACTOOL, we utilize a **Knowledge-based QA** approach. This also follows a QA format but incorporates the CoT method, where the LLM evaluates if *claims* are true or false relative to the source document. Details of the experiments are provided in Appendix B.

	AGGREFACT-CNN			AGGREFACT-XSUM			
	FTSOTA	EXF	OLD	FTSOTA	EXF	OLD	AVG
Baseline	50.0	50.0	50.0	50.0	50.0	50.0	50.0
DAE*	59.4	67.9	69.7	**73.1**	-	-	67.5
QuestEval	63.7	64.3	65.2	61.6	60.1	59.7	62.4
SummaC-ZS	63.3	**76.5**	76.3	56.1	51.4	53.3	62.8
SummaC-Cv	70.3	69.8	78.9	67.0	64.6	67.5	69.7
QAFactEval	61.6	69.1	**80.3**	65.9	59.6	60.5	66.2
AlignScore	53.4	73.1	80.2	70.2	**80.1**	63.7	70.1
ChatGPT-ZS	66.2	64.5	74.3	62.6	69.2	60.1	66.2
ChatGPT-CoT	49.7	60.4	66.7	56.0	60.9	50.1	57.3
ChatGPT-DA	48.0	63.6	71.0	53.6	65.6	61.5	60.6
ChatGPT-Star	55.8	65.8	71.2	57.7	70.6	53.8	62.5
FactScore	69.9	71.6	73.9	68.0	63.5	66.8	69.0
FacTool	72.7	66.1	60.8	68.0	64.0	62.2	65.6
FIZZ (Ours)	**73.2**	67.3	76.0	69.7	72.4	**68.5**	**71.2**

Table 2: Balanced accuracy on the AGGREFACT dataset. As in Tang et al. (2023), we omitted the results from DAE, as it was trained on the XSumFaith (Goyal and Durrett, 2021) dataset, which includes human-annotated summaries from EXFORMER and OLD.

4.4 Results

We present the performance outcomes obtained by applying each metric to the AGGREFACT benchmark dataset in Table 2. We show the performance of three versions of our proposed metric: FIZZ, its without granularity expanded version, FIZZ$_{w/o\ GE}$, and its without atomic facts version, FIZZ$_{w/o\ AF}$. The complete results for AGGREFACT-CNN and AGGREFACT-XSUM are displayed in Table 2. FIZZ demonstrates the highest average performance, followed by FIZZ$_{w/o\ GE}$ and FIZZ$_{w/o\ AF}$.

Additionally, we provide results for a single-threshold approach on AGGREFACT-FTSOTA split as in Tang et al. (2023). We list the best threshold findings for the AGGREFACT-CNN-FTSOTA and AGGREFACT-XSUM-FTSOTA splits, with corresponding binary classification balanced accuracy scores in Table 1. In this setting, FIZZ achieves the highest average performance, with FIZZ$_{w/o\ GE}$ coming in second. Both metrics perform exceptionally well on the CNN split. Furthermore, the granularity expansion in FIZZ leads to notably higher performance improvements on the XSUM split.

Both FACTSCORE and FACTOOL have demonstrate scores that are comparable to or exceed those of ChatGPT-based metrics. It appears that decomposing summaries into atomic facts and comparing them with the source document is more effective than performing factuality checking on the entire summary. However, metrics based on ChatGPT inherently face disadvantages compared to other metrics, which can be tuned by adjusting thresholds;

LLM	CNN	XSUM	AVG	AVG. TOKEN LENGTH
Zephyr	65.1±3.3	65.2±2.0	65.2	**97.6**
gpt-3.5-turbo	68.7±3.4	68.7±2.0	68.7	95.9
gpt-3.5-turbo-instruct	70.7±3.1	67.0±1.8	68.9	90.5
Mistral	70.5±3.5	68.7±2.1	69.6	86.5
Orca-2	**72.6**±3.0	**69.3**±1.9	**71.0**	81.4

Table 3: Experimental results of FIZZ with atomic facts generated by different LLMs using the same prompt on AGGREFACT-FTSOTA split. **Avg. Token Length** indicates the average number of total tokens of atomic facts per summary.

such tuning is unnecessary for ChatGPT-based metrics. This distinction may limit the effectiveness of ChatGPT-based evaluations in some contexts.

4.5 Analysis

LLMs used for Atomic Facts Decomposition To investigate the most suitable LLMs for generating atomic facts, we evaluate the generation of atomic facts using various LLMs, including `gpt-3.5-turbo`, `gpt-3.5-turbo-instruct`, and other 7B models such as Zephyr (Tunstall et al., 2023) and Mistral (Jiang et al., 2023). The results, documented in Table 3, demonstrate that while the atomic facts generated by `gpt-3.5-turbo` and `gpt-3.5-turbo-instruct` generally perform better compared to other metrics, they are still inferior to those produced by Orca-2. The performance drop associated with the `gpt` series suggests a noteworthy observation. We explain that this discrepancy is due to the length of the atomic facts. As shown in Table 3, which includes the average token length of atomic facts after the filtering process per summary, there is a clear inverse relationship between the number of tokens in an atomic fact and its average performance. Longer atomic facts tend to contain more entities and are less concise. Such sentences are less suitable as *hypotheses* when compared sentence-wise using NLI models. Furthermore, the sensitivity of using the minimum atomic fact scores as the final score exacerbates the challenge, making it difficult to achieve desired outcomes with lengthy sentences. In contrast, other 7B models such as LLaMa (Touvron et al., 2023) show limitations in adhering to instructions for atomic fact decomposition. Details of the model usage are provided in Appendix C.

	ROUGE-1			AVG. NUMBER OF ATOMIC FACTS	AVG. TOKEN LENGTH
	P	R	F1		
Human	1.00	1.00	1.00	8.7	98.4
Orca-2	0.70	0.69	0.68	**8.7**	**96.3**
gpt-3.5-turbo	**0.78**	**0.84**	**0.79**	7.8	105.0
gpt-3.5-turbo-instruct	0.73	0.72	0.70	13.0	149.6
Mistral	0.63	0.62	0.61	9.6	104.1
Zephyr	0.51	0.60	0.52	10.1	122.0

Table 4: Experimental results of generated atomic facts on RoSE dataset. The results with the highest human correlation are highlighted in **bold**.

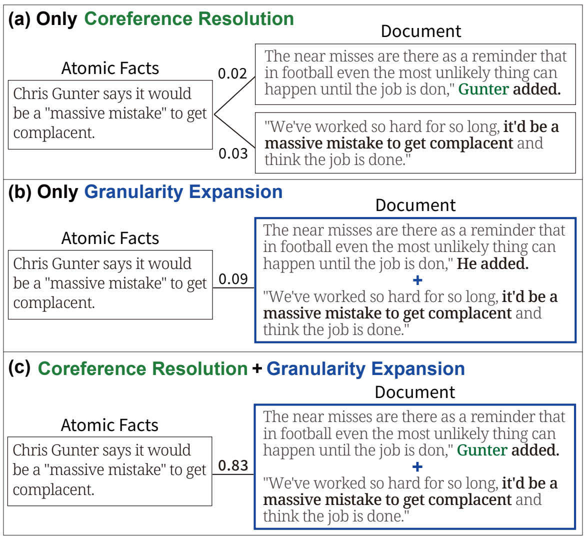

(a) Only Coreference Resolution
Document
Atomic Facts
Chris Gunter says it would be a "massive mistake" to get complacent.
0.02
0.03
The near misses are there as a reminder that in football even the most unlikely thing can happen until the job is don," Gunter added.
"We've worked so hard for so long, it'd be a massive mistake to get complacent and think the job is done."
(b) Only Granularity Expansion
Document
Atomic Facts
Chris Gunter says it would be a "massive mistake" to get complacent.
0.09
The near misses are there as a reminder that in football even the most unlikely thing can happen until the job is don," He added.
+
"We've worked so hard for so long, it'd be a massive mistake to get complacent and think the job is done."
(c) Coreference Resolution + Granularity Expansion
Document
Atomic Facts
Chris Gunter says it would be a "massive mistake" to get complacent.
0.83
The near misses are there as a reminder that in football even the most unlikely thing can happen until the job is don," Gunter added.
+
"We've worked so hard for so long, it'd be a massive mistake to get complacent and think the job is done."

Figure 3: The effect of granularity expansions and coreference resolution in real AGGREFACT dataset. The entailment score of an atomic fact and document sentence with (a) only Coreference Resolution, (b) only Granularity Expansion, and (c) the both.

In previous studies (Zhang and Bansal, 2021; Chern et al., 2023; Scirè et al., 2024), the evaluation of the quality and the completeness of the LLM generated atomic facts focuses solely on content similarity (*i.e.*, ROUGE-1) with human-written atomic facts. However, we consider content similarity evaluation to be insufficient and added two additional factors: 1) Average token length in atomic facts and 2) Average number of atomic facts. In Table 3, we demonstrate the correlation between the average token length of atomic facts and overall performance. Building on this, we now analyze the token length of both human-written and generated atomic facts. Additionally, since the content similarity metric does not take into account the number of atomic facts, we also include the average number of atomic facts in our results. We report the comparative analysis of the LLM generated atomic facts against human-written atomic facts in Table 4. The experiments were implemented using the RoSE (Liu et al., 2023) dataset, which includes 2,500 summaries and their corresponding human-written atomic facts. As shown in the experimental results, `gpt-3.5-turbo` demonstrates the highest capability by achieving the top score in content similarity. However, it shows a significant

Doc. Max Granularity	AGGREFACT-CNN-FTSOTA	AGGREFACT-XSUM-FTSOTA	AVG	s/it
One Sent.	72.2±2.8	66.3±1.9	69.25	2.49
Two Sent.	71.0±3.2	69.3±2.0	70.15	2.53
Three Sent.	**72.6**±3.0	69.3±1.9	70.95	2.64
Four Sent.	72.1±3.1	**70.0**±1.8	**71.05**	2.80

Table 5: **Size of granularity choice** in granularity expansion on AGGREFACT-FTSOTA split. s/it indicates seconds per iteration for the inference of an NLI model.

Atomic Facts	Doc	CNN	XSUM	AVG
Original	Original	63.2±2.3	66.4±1.8	64.8
	Coref Resolved	65.7±3.4	**67.8**±2.0	66.7(+*1.95*)
Coref Resolved	Original	66.2±3.4	66.6±1.9	66.4
	Coref Resolved	**72.2**±2.7	66.3±1.9	**69.2**(+*2.85*)

Table 6: **Effect of coreference resolution** of document and atomic facts on AGGREFACT-FTSOTA splits before the process of granularity expansion.

difference in the number of atomic facts and the number of tokens in atomic facts. In contrast, Mistral scores lower in content similarity but exhibits higher human correlation in the number of atomic facts and token lengths. The model that achieves the highest human correlation in both the number of atomic facts and token lengths is Orca-2, which shows the best performance among LLMs as in Table 3. These findings suggest that while content similarity is important, the number of atomic facts and token lengths are equally critical factors to consider. Details on computing content similarity are provided in Appendix G.

Sizes of Granularity Expansion As underscored in Section 3.3, accurately evaluating the factual consistency of *abstractive* summaries necessitates an expansion of document granularity. This requirement stems from the observation that a single sentence within a summary may incorporate content from multiple sentences within the document. Illustrative of this point, Figure 3 highlights that segmenting conversational dialogues into discrete sentences can lead to a loss of contextual clarity, where the synthesis of various segmented sentences is required for an accurate interpretation.

SUMMAC present experimental results across different granularity choices, categorizing document granularity into a sentence, two sentences, paragraph, and full document levels. However, adjusting document granularity in such a manner reduces interpretability and undermines result reliability. Our approach is to adaptively increase granularity only for atomic facts where the entailment score significantly decreases.

Table 5 presents the outcomes associated with varying granularity sizes in adaptive granularity expansion. The experimental findings reveal a consistent improvement in average performance with increasing granularity, particularly for summaries derived from XSum (Narayan et al., 2018). This significant performance boost can be attributed to the inherently *abstractive* nature of XSum-based summaries.

Despite the increase in average score for the maximum of four sentences, the scores for CNN summaries actually declined. Additionally, we observe that computational costs rose with increasing granularity. Hence, we determined that the maximum of three sentences represents the best tradeoff between computational cost and performance. Details on granularity expansion condition choice are provided in Appendix F.

Effectiveness of Coreference Resolution In the application of NLI models for comparing *premises* with *hypotheses*, the significance of coreference resolution cannot be overstated. As outlined in Section 3.1, failure to resolve pronouns in the *premise* significantly hinders the attainment of desired outcomes. This point is vividly illustrated in Figure 3, where the difference between document(b) and document(c) is merely the resolution of pronouns. Yet, this seemingly minor modification leads to a stark contrast in entailment scores, with document(b) achieving a score of 0.09 compared to document(c)'s 0.83. The discrepancy arises due to the document (*premise*)'s reference to "*he*" not being recognized as pertaining to "*Chris Gunter*", as stated in the atomic fact (*hypothesis*).

Moreover, Table 6 presents more granular experimental results on the impact of coreference resolution. We implemented experiments to evaluate the impact of coreference resolution on both documents and atomic facts. Our investigation included scenarios where coreference resolution was applied and cases where it was not. We show that texts with resolved coreferences, whether they be atomic facts or documents, consistently outperform those without resolution. Notably, there is a marked improvement in performance on datasets based on CNN (Hermann et al., 2015) summaries compared to those based on XSum summaries. This is likely due to the *extractive* nature of CNN-based summaries, as opposed to the more *abstractive* summaries derived from XSum. Details on coreference

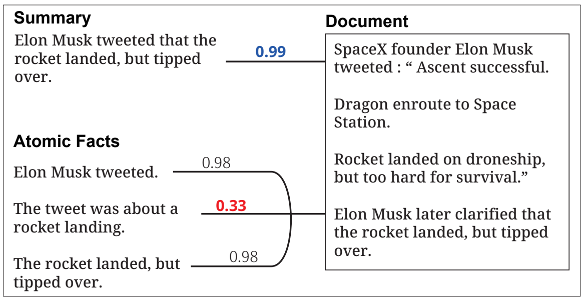

Summary
Elon Musk tweeted that the rocket landed, but tipped over.
0.99
Document
SpaceX founder Elon Musk tweeted : " Ascent successful.
Dragon enroute to Space Station.
Rocket landed on droneship, but too hard for survival."
Elon Musk later clarified that the rocket landed, but tipped over.
Atomic Facts
Elon Musk tweeted.
0.98
The tweet was about a rocket landing.
0.33
The rocket landed, but tipped over.
0.98

Figure 4: Drawbacks of atomic fact level evaluation versus the sentence level evaluation. The numbers represent the maximum NLI entailment scores obtained by comparing each sentence and atomic fact with the source document on a sentence-wise basis.

resolution usage are provided in Appendix E.

Failure Case Study We analyze the drawbacks of decomposing summaries into atomic facts in the summary factual consistency checking task, through the main example in Figure 4, which compares the drawbacks of analyzing atomic facts versus sentences. When comparisons are made at the sentence level, a sentence can be correctly judged as entailing the content of a document. Conversely, when breaking down the content into atomic facts, the fact *"The tweet was about a rocket landing."* receives a maximum entailment score of only 0.33. This particular atomic fact remains even after undergoing the filtering process. As a result, a summary that is factually consistent may be erroneously classified as factually inconsistent due to the analysis of this single atomic fact.

5 Conclusion

In this work, we propose a novel method, FIZZ, in detecting summary factual inconsistency. Our approach decomposes summaries into *atomic facts* and conducts a sentence-wise comparison with the document, and achieves state-of-the-art performance on the AGGREFACT benchmark dataset. Also, our proposed system has a higher interpretability due to its ability to precisely identify which parts of a summary are factually inaccurate by breaking it down into *atomic facts*. Furthermore, we analyze the necessity and significance of coreference resolution and granularity expansion in the context of summary factual consistency checking.

Limitations

Our proposed method is quite time-consuming. Notably, during the coreference resolution phase, we leverage 11B model. This process requires more time than other factual consistency checking systems. The practical applicability of FIZZ in real-time settings remains to be determined.

Furthermore, our research was limited to summaries based on articles and news domains. We did not verify the effectiveness of FIZZ in other domains such as dialogue summarization (Tang et al., 2024) or medical summarization (Wang et al., 2023b). Additionally, our study was confined to English-language data. The validity of FIZZ needs to be assessed in datasets based on other languages.

Despite these limitations, we believe our method paves a new path in the area of summarization factual consistency detection. This work could be a significant contribution to the field, pending further validation across varied domains and languages.

Ethics Statement

This work uses English document summarization dataset, AGGREFACT. This dataset is publicly available online. We also provided adequate citations for the papers and sources we consulted in writing our paper. Our work may have implications for society in terms of preventing the spread of inaccurate information, as it deals with factual consistency checking.

Acknowledgement

This research was supported by the Chung-Ang University Research Grants in 2023. This research was partly supported by Institute for Information & Communications Technology Planning & Evaluation (IITP) through the Korea government (MSIT) under Grant No. 2021-0-01341 (Artificial Intelligence Graduate School Program (Chung-Ang University)).

References

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Stephen Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Bernd Bohnet, Chris Alberti, and Michael Collins. 2023. Coreference resolution through a seq2seq transitionbased system. *Transactions of the Association for Computational Linguistics*, 11:212–226.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models.

Shiqi Chen, Siyang Gao, and Junxian He. 2023. Evaluating factual consistency of summaries with large language models.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

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.

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.

Yanjun Gao, Chen Sun, and Rebecca J. Passonneau. 2019. Automated pyramid summarization evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 404–418, Hong Kong, China. Association for Computational Linguistics.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. TrueTeacher: Learning factual consistency evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2053–2070, Singapore. Association for Computational Linguistics.

John Glover, Federico Fancellu, Vasudevan Jagannathan, Matthew R. Gormley, and Thomas Schaaf. 2022. Revisiting text decomposition methods for NLI-based factuality scoring of summaries. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 97–105, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 166–175, New York, NY, USA. Association for Computing Machinery.

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

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.

Aaron Harnly, Ani Nenkova, Rebecca Passonneau, and Owen Rambow. 2005. Automation of summary evaluation by the pyramid method. In *International Conference on Recent Advances in Natural Language Processing, RANLP 2005 - Proceedings*, International Conference Recent Advances in Natural Language Processing, RANLP, pages 226–232. Association for Computational Linguistics (ACL). International Conference on Recent Advances in Natural Language Processing, RANLP 2005 ; Conference date: 21-09-2005 Through 23-09-2005.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

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

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

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.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

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

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.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching small language models how to reason.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. 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.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

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

OpenAI. 2022. Chatgpt blog post. `https://openai.com/blog/chatgpt`.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Tal Schuster, Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, and Donald Metzler. 2022. Stretching sentence-pair NLI models to reason over long documents and clusters. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 394–412, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, 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.

Alessandro Scirè, Karim Ghonim, and Roberto Navigli. 2024. FENICE: Factuality evaluation of summarization based on natural language inference and claim extraction. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14148–14161, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Ori Shapira, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2019. Crowdsourcing lightweight pyramids for manual summary evaluation. 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 682–687, Minneapolis, Minnesota. Association for Computational Linguistics.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the factual consistency of large language models through news summarization. In *Findings of* *the Association for Computational Linguistics: ACL 2023*, pages 5220–5255, Toronto, Canada. Association for Computational Linguistics.

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.

Liyan Tang, Igor Shalyminov, Amy Wing mei Wong, Jon Burnsky, Jake W. Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, Yi Zhang, Saab Mansour, and Kathleen McKeown. 2024. Tofueval: Evaluating hallucinations of llms on topic-focused dialogue summarization.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Hans van Halteren and Simone Teufel. 2003. Examining the consensus between human summaries: initial experiments with factoid analysis. In *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*, pages 57–64.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Lucy Lu Wang, Yulia Otmakhova, Jay DeYoung, Thinh Hung Truong, Bailey Kuehl, Erin Bransom, and Byron Wallace. 2023b. Automated metrics for medical multi-document summarization disagree with human evaluations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9871–9889, Toronto, Canada. 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, New Orleans, Louisiana. Association for Computational Linguistics.

Jiuding Yang, Hui Liu, Weidong Guo, Zhuwei Rao, Yu Xu, and Di Niu. 2024. Sifid: Reassess summary factual inconsistency detection with llm.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Shiyue Zhang and Mohit Bansal. 2021. Finding a balanced degree of automation for summary evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6617–6632, Online and 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.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *ArXiv*, abs/2309.01219.

A Prompt for Atomic Facts Decomposition

The prompt for atomic fact decomposition in shown in Table 10. The examples given in the prompt are similarly used in other LLMs.

B Details on Baselines

In this section, we present the implementation details of FActScore and FacTool, which have been integrated into our experimental baseline. For decomposing atomic facts, FActScore uses the `gpt-3.5-turbo-instruct` model, and the QA process is conducted using `gpt-3.5-turbo`, with prompts exactly as specified in the paper[2]. We gave 1 point for each answer that is answered ture and then divided by the total number of atomic facts:

$$score = \frac{1}{|A|} \sum_{a \in A} \mathbb{I}[\text{ a is True }] \tag{4}$$

Similar to FActScore, FacTool employs `gpt-3.5-turbo` for both the *claim* extraction and the QA tasks, again using prompt directly from the paper[3].

C Details on the Usage of Large Language Models

We report on the details and `Huggingface` links of LLMs used in Section 4. We employed Orca-2-7B model[4] for experiments in AggreFact benchmark dataset. For Zephyr, we used Zephyr-7B-beta [5], while for Mistral, we used Mistral-7B-instruct-v0.2 [6]. Additionally, we used ChatGPT version of `gpt-3.5-turbo-0125`.

D Details on the Usage of NLI Model

In this study, we tried to analyze the effect of our proposed atomic fact level decomposition instead of using entire sentences. To ensure a fair comparison of our approach with SummaC, which demonstrated the best performance using whole sentences, we employed the same NLI model that was utilized in SummaC[7]. The model has been trained on the conventional NLI datasets SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), ANLI (Nie et al., 2020), and also on VitaminC (Schuster et al., 2021).

In Table 7, we present the performance results of various NLI models. Specifically, we have included the results for `DeBERTa-large-mnli`[8] and `RoBERTa-large-pyrxsum`[9]. The average performance scores for `DeBERTa` and `RoBERTa` are 68.7 and 68.5, respectively. Although these scores are lower than that of `ALBERT`, they surpass the previous best score of 67.8 achieved by DAE on the FtSota split.

NLI Model	AggreFact-CNN-FtSota	AggreFact-XSum-FtSota	Avg
ALBERT	**72.6**±3.0	69.3±1.9	**71.0**
DeBERTa	67.3±3.0	**70.1**±1.9	68.7
RoBERTa	70.5±3.0	66.5±1.9	68.5

Table 7: Performance of different NLI models on AggreFact-FtSota split.

E Details on the Usage of Coreference Resolution

We used MT5-11B model for coreference resolution[10]. Coreference resolution is the task of identifying all expressions that refer to the same entity within a text. While recent models perform well on this task, returning a text with resolved coreferences is an entirely different challenge. We have tested various models, but none have functioned adequately. A significant issue was the prevalent method of using the first word in a cluster for resolution instead of the entity's name, which frequently resulted in improper replacements with pronouns. To address this, we slightly modified the code to ensure that where an entity name is available, it replaces pronouns as much as possible[11]. Furthermore, when an adjective or a modifier refers to an entity, we prefixed it with the entity's name followed by a comma. Table 11 illustrates these modifications. By enhancing coreference resolution in this manner, we were able to capture

[2] https://github.com/shmsw25/FActScore

[3] https://github.com/GAIR-NLP/factool

[4] `https://huggingface.co/microsoft/Orca-2-7b`

[5] `https://huggingface.co/HuggingFaceH4/zephyr-7b-beta`

[6] `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2`

[7] `https://huggingface.co/tals/albert-xlarge-vitaminc-mnli`

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

[9] `https://huggingface.co/shiyue/roberta-large-pyrxsum`

[10] `https://huggingface.co/mt5-coref-pytorch/link-append-xxl`

[11] `https://github.com/google-research/google-research/tree/master/coref_mt5`

Condition	AggreFact-CNN-FtSota	AggreFact-XSum-FtSota	Avg
!(e>c & e>n)	**72.6**±3.0	**69.3**±1.9	**71.0**
!(e>c || e>n)	71.1±2.9	68.7±1.9	69.9

Table 8: Granularity Expansion condition choice on AggreFact-FtSota split.

more comprehensive atomic facts without omitting critical information.

F Details on Granularity Expansion

In Section 3.3, we set the criterion for granularity expansion as $\max(e, c, n)! = e$. This criterion was chosen because it intuitively signifies a lack of entailment. Notably, $\max(e, c, n)! = e$ is equivalent to $!(e > c \,\&\, e > n)$, and thus, we also conducted experiments using the $!(e > c \,||\, e > n)$ condition. Table 8 presents the results of these experiments.

G Details on Computing Content Similarity

The content similarity (ROUGE-1) in Table 4 was conducted using the following equation:

$$\frac{1}{N_{data}} \sum_{N_{data}} \frac{1}{N_c} \sum_{i=1}^{N_c} \max_{j=1}^{N_g} (\text{ROUGE}(c_i, g_j)) \quad (5)$$

where c denotes LLM generated atomic facts and g denotes human-written atomic facts.

H Other Details

In this section, we report the differences observed when splitting text into sentences using NLTK (Bird et al., 2009) and Spacy (Honnibal et al., 2020). We utilized NLTK sentence splitter in FIZZ. The results of the experiments are presented in Table 9.

Sentence Splitter	AggreFact-CNN-FtSota	AggreFact-XSum-FtSota	Avg
Spacy	72.5±3.4	67.0±2.0	69.8
NLTK	**72.6**±3.0	**69.3**±1.9	**71.0**

Table 9: Sentence splitter choice on AggreFact-FtSota split.

Input Prompt

You are a helpful assistant. Please give me a list of atomic facts of the following texts.

lisa courtney, of hertfordshire, has spent most of her life collecting pokemon memorabilia.
- Lisa Courtney is from Hertfordshire.
- Lisa Courtney has spent most of her life collecting Pokémon memorabilia.

prince jan zylinski said he was fed up with discrimination against poles living in britain.
- Prince Jan Zylinski made a statement.
- The statement made by Prince Jan Zylinski was about discrimination.
- The statement made by Prince Jan Zylinski was regarding Poles living in Britain.
- Prince Jan Zylinski expressed feeling fed up with this type of discrimination.

no charges were filed, there will be no travel ban.
- No charges were filed.
- There will be no travel ban.

rudd has pleaded guilty to threatening to kill and possession of drugs in a court.
- Rudd has pleaded guilty.
- Rudd has pleaded guilty to threatening to kill.
- Rudd has pleaded guilty to possession of drugs.

Lee made his acting debut in the film The Moon is the Sun’s Dream (1992), and continued to appear in small and supporting roles throughout the 1990s.
- Lee made his acting debut in The Moon is the Sun’s Dream.
- The Moon is the Sun’s Dream is a film.
- The Moon is the Sun’s Dream was released in 1992.
- After Lee’s acting debut, he appeared in small and supporting roles throughout the 1990s.

In 1963, Collins became one of the third group of astronauts selected by NASA and he served as the back-up Command Module Pilot for the Gemini 7 mission.
- Collins became an astronaut.
- Collins became one of the third group of astronauts selected by NASA in 1963.
- Collins served as the back-up Command Module Pilot for the Gemini 7 mission.

In addition to his acting roles, Bateman has written and directed two short films and is currently in development on his feature debut.
- Bateman has acting roles.
- Bateman has written two short films.
- Bateman has directed two short films.
- Bateman is currently in development on his feature debut.

Michael Collins (born October 31, 1930) is a retired American astronaut and test pilot who was the Command Module Pilot for the Apollo 11 mission in 1969.
- Michael Collins was born on October 31, 1930.
- Michael Collins is retired.
- Michael Collins is an American.
- Michael Collins was an astronaut.
- Michael Collins was a test pilot.
- Michael Collins was the Command Module Pilot for the Apollo 11 mission in 1969.

Summary Sentence

Table 10: Prompt used to generate atomic facts from coreference resolved summary in Section 3.2. We employed 8-shot learning to enhance the model’s performance.

Original Text		**The 27-year-old** joined spurs from manchester city in 2011.
Others	**Coref Resolved Text**	**Emmanuel Adebayor** joined spurs from manchester city in 2011.
	Atomic Fact #1	Emmanuel Adebayor joined spurs.
	Atomic Fact #2	Emmanuel Adebayor joined spurs from manchester city.
	Atomic Fact #3	Emmanuel Adebayor joined spurs in 2011.
Ours	**Coref Resolved Text**	**Emmanuel Adebayor, the 27-year-old** joined spurs from manchester city in 2011.
	Atomic Fact #1	**Emmanuel Adebayor is 27-year-old.**
	Atomic Fact #2	Emmanuel Adebayor joined spurs.
	Atomic Fact #3	Emmanuel Adebayor joined spurs from manchester city.
	Atomic Fact #4	Emmanuel Adebayor joined spurs in 2011.

Table 11: Our distinct approach for coreference resolution. The original text is coreference resolved by two ways, which are **Others** and **Ours**. We ensure that critical information is preserved while generating atomic facts by prefixing modifiers with the names of entities during the coreference resolution.