Granularity is the Hidden Bias: Revisiting the Factual Consistency Evaluation for Summarization

1st Weina Ma
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
maweina@iie.ac.cn

2nd Jiayang Gao
Beijing Institute of Computer Technology and Application
Beijing, China
nemo830218@gmail.com

3rd Zimo Nie
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
niezimo@iie.ac.cn

4rd Zhiwei Yang
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
yangzhiwei@iie.ac.cn

5th Chaodong Tong*
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
tongchaodong@iie.ac.cn

6th Chengqing Guo*
National Computer Network Emergency Response Technical Team/Coordination Center of China
Beijing, China
chqguo2006@gmail.com

7th Lei Jiang
Institute of Information Engineering
Chinese Academy of Sciences
Beijing, China
jianglei@iie.ac.cn

***Abstract*—Factual consistency evaluation for summarization has attracted increasing attention as large language models (LLMs) increasingly generate hallucinations. Existing evaluation methods implicitly obey the fixed-fact-granularity assumation (verifying summaries under a uniform claim segmentation granularity) implicitly when decomposing and verifying summary content. This assumption inevitably introduces a hidden but systematic evaluation bias in real-world scenarios: logically independent facts may be merged to mask errors, while semantically interdependent complex facts may be over-splitting and break relational semantics. We revisit this problem from the perspective of granularity control, revealing that granularity mismatch is a fundamental failure mode that affects the reliability and interpretability of the evaluation. To address this problem, we propose a factual consistency evaluation framework FAIR, which achieves explicit modeling of factual granularity by performing logically aware fact segmentation on the summary and semantically aware evidence segmentation on the source documents. Furthermore, FAIR employs a retrieval-enhanced verification mechanism to perform precise alignment and verification of atomic content units and outputs concrete error types. Experiments on benchmark datasets demonstrate that FAIR achieves stronger correlation with human fidelity judgments and reduces evaluation bias, delivering both interpretable and reliable assessment of LLM-based summarization.**

***Index Terms*—Factual Consistency Evaluation, Evaluation Bias, Granularity Control, Retrieval-Enhanced Verification**

I. Introduction

Factual consistency evaluation for summarization has attracted growing attention as large language models (LLMs) increasingly generate hallucinations (statements that are fluent and plausible yet unsupported by the source document) [1], [2]. Factual consistency evaluation aims to determine whether each factual content expressed in a summary is grounded in its source [3], ideally that can localize errors and provide actionable feedback.

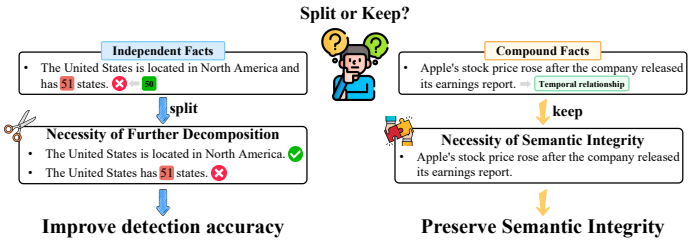

Fig. 1. Illustration of the motivation for logic-aware segmentation. Independent facts (left) are split into smaller atomic units to reveal errors and improve detection accuracy, whereas compound facts (right) need to be preserved as integral units to maintain semantic integrity.

However, existing evaluation methods commonly follow a fixed-fact-granularity assumption, i.e., verifying summaries under a uniform claim segmentation granularity, when decomposing and checking summary content [4], [5], [6]. This implicit assumption becomes problematic in practice because factual statements in summaries are not homogeneous (as shown in Fig. 1): independent facts should be logically separated for precise error localization (standalone propositions within one sentence), whereas compound facts are semanti-

* denotes the corresponding author.

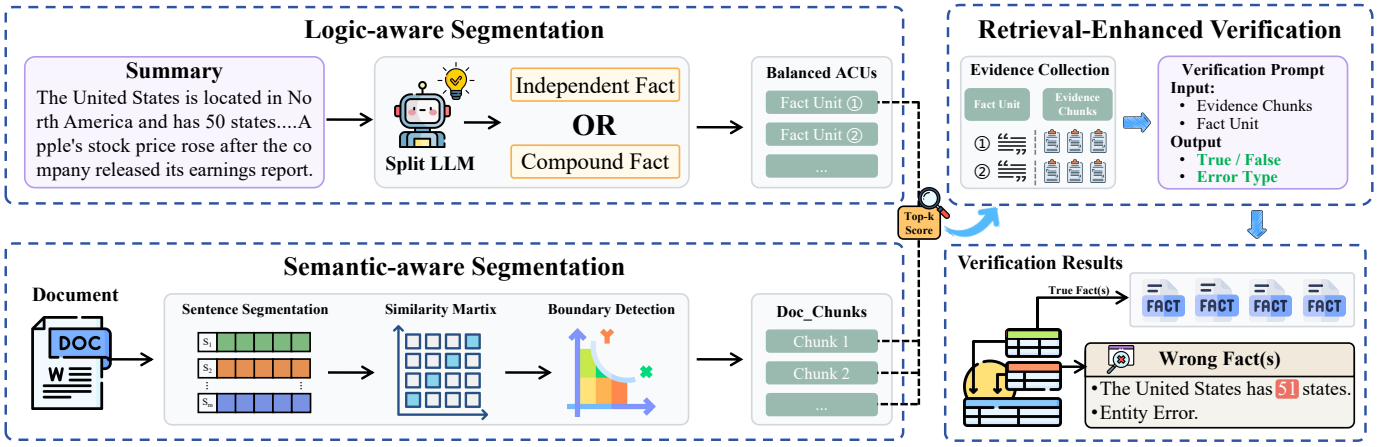

Fig. 2. Pipeline of FAIR. Given a document-summary pair, the logic-aware segmentation module decomposes the summary into granularity-controlled atomic content units (ACUs), whereas the semantic-aware segmentation module segments the source document into semantically coherent chunks. The evidence collection module semantically aligns each ACU with the top-K document chunks, and the verifier synchronously outputs the authenticity (true/false) and its corresponding error type.

cally interdependent and rely on internal relational semantics (e.g., temporal or causal links) that must be preserved. As a consequence, a hidden but systematic evaluation bias emerges: logically independent facts may be merged to mask errors, while semantically interdependent complex facts may be over-splitting and breaking relational semantics. Such granularity mismatch-induced failures undermine both the reliability of metric scores and the interpretability of the resulting evaluation. Therefore, we argue that granularity mismatch is a fundamental failure mode in the factual consistency evaluation task.

To address this problem, we revisit factual consistency evaluation from the perspective of granularity control and propose **FAIR** (**F**actual **A**ssessment with **I**nterpretable and **R**eliable Framework), as shown in Fig. 2, which explicitly models factual granularity on both sides of the verification process. Specifically, FAIR performs logic-aware fact segmentation on the summary to obtain granularity-controlled atomic content units (ACUs) and conducts semantic-aware evidence segmentation on the source document to construct semantically coherent evidence chunks [7], [8]. Furthermore, FAIR employs a retrieval-enhanced verification mechanism to perform precise alignment and verification of atomic content units and output concrete error types [9], [10]. This design enables FAIR to provide not only an overall factual consistency score but also interpretable ACU-level verification records that expose where and how a summary becomes unfaithful.

We evaluate FAIR on multiple benchmark datasets for summarization factuality evaluation. Experimental results demonstrate that FAIR achieves stronger correlation with human fidelity judgments and reduces hidden bias in evaluation compared with representative baselines [11], [12] while simultaneously producing interpretable diagnostic outputs. These findings suggest that explicitly granularity-controlling evaluation is essential for building reliable and interpretable factual consistency evaluation for summarization in the LLM era.

Our contributions can be summarized as follows:

- To the best of our knowledge, this is the first work to investigate the hidden bias in factual consistency evaluation through the lens of granularity control.
- We propose FAIR, a granularity-controlling evaluation framework that jointly performs logic-aware and semantic-aware segmentation, enabling retrieval-enhanced ACU verification with explicit error typing.
- Extensive experiments on three benchmarks show that FAIR improves alignment with human judgments, reduces evaluation bias, and achieves competitive or state-of-the-art performance baselines.

II. Related Work

Factual consistency evaluation for summarization has received increasing attention with the rise of LLM-based generation, where fluent but ungrounded statements are common. From the granularity-controlling perspective, existing work can be organized by the evaluation unit it verifies—global-level, sentence-level, and claim-level factuality evaluation. This granularity choice is not merely an implementation detail: it implicitly defines what constitutes a ”checkable fact” and can introduce systematic bias when the chosen unit mismatches the underlying discourse structure.

Global-level factuality evaluation. Global-level factuality evaluation treats an entire summary (or large spans) as a single unit and produces a holistic score via reference- or source-based matching. Classic methods such as ROUGE [13] and embedding-based similarity methods like BERTScore [14] and MoverScore [15] are widely used due to their simplicity and efficiency. However, from a granularity perspective, such holistic scoring is inherently mismatched with factual verification: by collapsing multiple propositions into a single comparison, it cannot isolate which specific statement is unsupported, and

hallucinated content may remain lexically or semantically plausible under global matching, leading to weak penalties and limited diagnostic value. These limitations have been repeatedly observed in meta-evaluations and benchmarking studies of factual factuality metrics for summarization [3], [12]. As a result, global-level metrics provide limited interpretability for hallucination analysis, motivating factuality evaluators that operate on smaller, verifiable units.

Sentence-level factuality evaluation. Sentence-level factuality evaluation verifies each summary sentence (or clause) against the source document, typically through entailment-based or QA-based formulations. Representative NLI-based approaches model factuality as sentence-level entailment or inconsistency detection, such as FactCC [16] and SummaC [17], with structural variants further incorporating syntactic or dependency-level entailment for more precise judgments [18]. QA-based evaluators, including QAFactEval [19] and QuestEval [20], improve coverage by converting summary content into questions and validating answers against the source, supported by datasets and annotation efforts that characterize fine-grained factuality phenomena [21]. However, from a granularity perspective, sentence-level evaluation still implicitly assumes a fixed verification unit: a single sentence may bundle multiple logically independent propositions (masking localized errors) or express discourse-compositional meaning whose relational semantics (e.g., temporal/causal links) can be broken by naive splitting, leading to unstable or biased judgments. These limitations are particularly evident for LLM-generated summaries, where sentence-level units frequently exhibit granularity mismatch, leading to biased or unstable verification outcomes under fixed-unit evaluation [1], [22].

Claim-level factuality evaluation. Claim-level factuality evaluation decomposes summaries into smaller claims or atomic facts and verifies each unit against supporting evidence from the source document. Representative frameworks operationalize this paradigm with explicit claim extraction and evidence-based verification, such as FactScore for atomic fact checking in long-form generation [4], and summarization-oriented pipelines including FENICE [5] and FIZZ [23] that couple fine-grained decomposition with evidence alignment. Recent extensions further enhance claim-level evaluation with actionable feedback and stronger verifiers, e.g., ACUEval for fine-grained hallucination evaluation and correction [6], LLM-driven QA evaluation such as SummEQuAL [9], and multi-step inference to strengthen verification under complex evidence conditions [10]. However, from a granularity perspective, claim-level evaluation faces a central dilemma: overly aggressive decomposition may yield fragments that are no longer self-contained or verifiable, while insufficient decomposition can still merge independent facts and obscure localized errors. This sensitivity has been highlighted in analyses of claim decomposition, which show that decomposition quality and granularity choices substantially affect downstream verification outcomes [24]. Moreover, even with well-formed claims, document-side evidence is often segmented with fixed units (e.g., sentences, paragraphs, or windows), which can mix topics or repeated entities and introduce retrieval noise, motivating retrieval-enhanced verification [25] and semantic chunking for more coherent evidence units [8]. As a result, effective claim-level factuality evaluation requires granularity-controlled atomic content units and coherent evidence units, rather than simply adopting finer segmentation.

In summary, global-level evaluation methods [13], [14], [15] are efficient but weakly diagnostic; sentence-level valuation methods [16], [17], [18], [19], [20] strengthen grounding but still assume a fixed unit; claim-level valuation methods [4], [5], [6], [9], [23] improve localization yet face decomposition and evidence-unit mismatch challenges [8], [24]. Our work differs by explicitly framing granularity mismatch as a hidden but systematic source of evaluation bias and addressing it via granularity control on both sides of verification: logic-aware segmentation constructs granularity-controlled atomic content units on the summary side, semantic-aware segmentation forms coherent evidence chunks on the document side, and retrieval-enhanced verification aligns and checks each unit while outputting fine-grained error types for interpretability.

III. Methodology

A. Task Definition

Given a document-summary pair $M = (D, S)$, factual consistency evaluation aims to determine whether the information conveyed by the summary is supported by the source document. We view the summary S as a set of factual statements, which can be further decomposed into a set of atomic content units (ACUs) $U = \{a_1, \ldots, a_{|U|}\}$. In parallel, the document D is segmented into semantically coherent chunks $C = \{c_1, \ldots, c_{|C|}\}$ that serve as candidate evidence. For each ACU a_i, the evaluator retrieves a small evidence set $R(a_i) \subseteq C$ and outputs a binary factuality label $y_i \in \{0, 1\}$ indicating whether a_i is supported by at least one retrieved chunk. In addition to the overall factuality score aggregated over U, FAIR outputs error-type labels for unsupported ACUs to improve interpretability.

A key challenge is that the granularity of fact unit decomposition critically affects verification reliability. Under-segmentation may merge multiple independent facts and conceal localized errors, while over-segmentation may break a logically coherent composite fact into incoherent fragments, introducing artificial mismatches. Therefore, our goal is to construct granularity-controlled verification units—small enough to isolate independent factual claims yet coherent enough to preserve the semantics of compound statements—before performing retrieval and verification.

B. Logic-aware Segmentation Module

To preserve the discourse structure required for reliable factual verification, we introduce a logic-aware segmentation module that decomposes summaries into granularity-controlled atomic content units (ACUs). Following prior work [7], ACUs are defined as elementary information units that no longer need to be further split to reduce ambiguity in human

TABLE I
FACTUAL ERROR TAXONOMY FOR CONSISTENCY VERIFICATION.

	Category	Description	Example
PredE	Relation Error	The predicate in the summary statement is inconsistent with the source article.	The committee approved the proposal on Tuesday.
EntE	Entity Error	The primary arguments (or their attributes) of the predicate are wrong.	The report said Microsoft acquired the startup for $2 billion.
CircE	Circumstance Error	The additional information (e.g., location or time) specifying the circumstance around a predicate is wrong.	The earthquake struck on Friday in southern Turkey.
CorefE	Coreference Error	A pronoun/reference has a wrong or non-existing antecedent.	Apple released its quarterly results. They said the CEO would resign.
LinkE	Discourse Link Error	Error in how multiple statements are linked in discourse (e.g., temporal ordering/causal link).	The airline cancelled the flight because of heavy fog; later, the storm arrived.
OutE	Out of Article Error	The statement contains information not present in the source article.	The minister also announced a nationwide tax cut.
GramE	Grammatical Error	The grammar is so wrong that the sentence becomes meaningless.	The he vaccine approved has already start to be.

evaluation. Building on this notion, we introduce granularity-controlled ACUs that explicitly mitigate both under- and over-segmentation for factual verification: each unit is small enough to isolate independent facts, yet remains logically coherent when a statement relies on discourse relations.

Sentence-level logical labeling. Given a summary $S = \{s_1, s_2, \ldots, s_n\}$, we assign each sentence s_i a core logical label $\mathcal{L}(s_i)$ over a predefined discourse-relation set $\mathcal{R}$. Specifically,

$$\mathcal{L}(s_i) = \arg\max_{r \in \mathcal{R}} f_\theta(s_i, r), \tag{1}$$

where $f_\theta(s_i, r)$ outputs a compatibility score measuring how well relation r explains the discourse structure expressed in s_i.

Token-Level Logical Annotation. After obtaining $\mathcal{L}(s_i)$, we perform token-level logical annotation within s_i. For each token, we predict a local logical label and compute a matching score that quantifies its consistency with $\mathcal{L}(s_i)$.

Boundary Insertion Rules. We introduce an intra-sentence segmentation boundary when either: (1) the predicted logical relation between two adjacent tokens changes, indicating a local shift in discourse function; or (2) the matching score between a token's local label and $\mathcal{L}(s_i)$ falls below 0.3, a suggesting deviation from the dominant logical structure of the sentence. Applying these rules yields spans that are semantically compact yet logically coherent, which we treat as ACUs.

Logical Relation Inventory. The relation set $\mathcal{R}$ contains six categories: *causal*, *contrastive*, *entailment*, *conditional*, *parallel*, and *temporal*. This inventory enables fine-grained modeling of intra-sentence logical structures while keeping the segmentation process interpretable.

Resulting ACUs. Intuitively, a sentence may encode multiple independent claims linked by coordination or enumeration, where splitting improves error localization, or a compositional claim whose meaning hinges on explicit discourse links (e.g., temporal/causal/contrastive cues), where aggressive splitting would destroy verifiable semantics. By segmenting based on label transitions and label-core mismatches, our procedure directly targets the failure cases in Fig. 2, mitigating under-segmentation (merging multiple facts and masking localized inconsistencies) and over-segmentation (fragmenting a coherent fact into unverifiable pieces). The output ACU set U is then used for downstream alignment and verification, where each $a_i \in U$ will be assigned a factuality label $y_i \in \{0, 1\}$ based on retrieved document evidence.

Algorithmic summary. Overall, the module converts each sentence into token-level annotations, detects intra-sentence logical boundaries, and outputs the ACU set U. These ACUs serve as the basic units for downstream alignment and verification, where each $a_i \in U$ is later assigned a factuality label $y_i \in \{0, 1\}$ based on retrieved document evidence.

C. Semantic-aware Segmentation Module

Logic-aware splitting focuses on the summary side by producing structurally meaningful ACUs. To complement it, we further design a semantic-aware segmentation module on the document side to generate semantically coherent evidence chunks, which is crucial for robust retrieval and accurate matching.

Complementary to logic-aware splitting on the summary side, we segment the source document D into semantically coherent chunks that serve as evidence candidates. The goal is to ensure that each chunk concentrates on a relatively consistent topic or semantic region, thereby improving retrieval precision and reducing noise during verification. For a candidate span u in the document, we measure intra-span semantic cohesion as

$$\mathcal{S}(u) = \frac{1}{|u|^2} \sum_{i,j \in u} \cos\big(\phi(w_i), \phi(w_j)\big), \tag{2}$$

where $\phi(w)$ denotes the contextual embedding of token w. Intuitively, a high $S(u)$ indicates that tokens within the span are mutually consistent and likely describe a coherent piece of evidence, whereas a low $S(u)$ suggests topic drift or mixed content. We iteratively split spans until

$$\mathcal{S}(u) \geq \tau, \tag{3}$$

where τ is a semantic threshold. This procedure produces document chunks C that are compact enough for high-precision retrieval while retaining sufficient context for downstream verification, forming a stable evidence pool for the retrieval-enhanced module.

D. Retrieval-Enhanced Verification Module

After constructing granularity-controlled ACUs U and semantically coherent document chunks C, we verify each ACU $a_i \in U$ against the source evidence. The proposed retrieval-enhanced verification module operates in two stages: (i) evidence retrieval with refinement and (ii) factuality verification with explicit error typing.

Evidence retrieval and refinement. For each ACU a_i, we compute its semantic similarity to every document chunk $c_p \in C$ and select the top-K chunks as a compact evidence set:

$$R(a_i) = \{\, c_p \in C \mid \mathrm{rank}(\mathrm{sim}(a_i, c_p)) \leq K \,\}, \qquad (4)$$

where $\mathrm{sim}(a_i, c_p)$ denotes the similarity between a_i and c_p, and $\mathrm{rank}(\cdot)$ returns the position of c_p in the descending similarity order. Restricting evidence to top-K chunks constrains verification to a compact context window, which improves efficiency and reduces distraction from irrelevant document regions.

Although top-K retrieval provides a strong candidate set, it may still include weakly related chunks, particularly in documents with repeated entities or loosely connected topics. We therefore refine $R(a_i)$ using a similarity threshold τ: chunks with $\mathrm{sim}(a_i, c_p) < \tau$ are discarded, yielding a concise evidence set dominated by semantically aligned content. This filtering step reduces retrieval noise and improves verification robustness.

Consistency verification with explicit error types. Given the refined evidence set $R(a_i)$, a verifier determines whether a_i is factually supported and outputs a binary label $y_i \in \{0, 1\}$. For unsupported units, the verifier additionally assigns an explicit error type to make the evaluation interpretable and actionable. As summarized in Table I, we adopt seven categories: `PredE` (Relation Error), `EntE` (Entity Error), `CircE` (Circumstance Error), `CorefE` (Coreference Error), `LinkE` (Discourse Link Error), `OutE` (Out-of-Article Error), and `GramE` (Grammatical Error). This structured output goes beyond a single scalar score by enabling ACU-level diagnosis of where and why factual inconsistencies arise.

E. Overall Framework

FAIR evaluates factual consistency through a three-stage pipeline: (1) logic-aware segmentation decomposes the summary into granularity-controlled ACUs U; (2) semantic-aware segmentation partitions the source document into semantically coherent chunks C; and (3) retrieval-enhanced verification aligns each $a_i \in U$ with its most relevant evidence $R(a_i) \subseteq C$ to predict unit-level factuality and assigns an error type for unsupported units.

Metric Aggregation. For each ACU a_i, the verifier outputs a factuality decision $s_i \in \{0, 1\}$ conditioned on $R(a_i)$.

System Prompt.
You are a factual consistency verifier. Given one atomic content unit (ACU) and top-K evidence chunks extracted from the source document, decide whether the ACU is *explicitly* supported by the evidence.

Rules.
- Use only the provided evidence chunks; do not use external knowledge.
- If the evidence does not explicitly support the ACU, set $label = 0$.
- Output json only with exactly two keys: `label` and `error_type`. If $label = 1$, set $error\_type = null$; if $label = 0$, choose exactly one type from {PredE, EntE, CircE, CorefE, LinkE, OutE, GramE}.

Error Type Guidance (most salient).
- `OutE`: ACU content not stated in any evidence chunk.
- `EntE`: wrong entity/attribute/numeric value.
- `PredE`: wrong predicate/relation.
- `CircE`: wrong circumstance (e.g., time/location/quantity).
- `CorefE`: incorrect or missing antecedent.
- `LinkE`: incorrect temporal/causal discourse link.
- `GramE`: grammar makes the ACU unclear and unverifiable.

Input.
{`ACU`}
Evidence.
{`evidence_1`}, {`evidence_2`}, ...
Output (JSON only).
`{"label": 0/1, "error_type": type/null}`

Fig. 3. Prompt template used for retrieval-enhanced ACU verification.

The overall FAIR score is computed by averaging unit-level decisions:

$$\text{FAIR\_Score} = \frac{1}{|U|} \sum_{i=1}^{|U|} s_i, \qquad (5)$$

where $|U|$ is the number of ACUs. This aggregation is well-suited under granularity control: since each unit is designed to be semantically compact yet logically complete, the resulting score reflects the proportion of supported factual units, rather than being dominated by under-/over-segmentation artifacts.

End-to-End Outputs. Beyond the scalar score in Eq. (5), FAIR produces an interpretable verification record for each unit, including the retrieved evidence $R(a_i)$, the binary verdict s_i, and (when applicable) the error category defined in Table I. This structured output enables fine-grained diagnosis of factual failure patterns and supports transparent, bias-aware comparison across summarizers and datasets; see Table II for an example.

IV. EXPERIMENTS

A. Datasets

We conduct experiments on three widely used factuality benchmarks for summarization. SummEval [12] provides human annotations for summaries generated by both extractive and abstractive systems and is commonly used to assess alignment between automatic metrics and human factuality judgments. AggreFact [11] aggregates multiple factuality datasets;

TABLE II
ILLUSTRATIVE ACU-LEVEL VERIFICATION RECORD PRODUCED BY FAIR (ALIGNED WITH FIG. 2).

Inputs	**Document (excerpt):**	The United States is located in North America and has **50** states.…
	Summary (excerpt):	The United States is located in North America and has **51** states.
Intermediate Outputs	**Granularity-controlled ACUs (excerpt):**	a_1: The United States is located in North America. a_2: The United States has **51** states.
	Document Chunks (excerpt):	$C = \{c_1, c_2, \ldots\}$, where c_1 contains: …has **50** states …
	Retrieved evidence (Top-3):	$R(a_1) = \{c_1, c_2, c_3\},\ R(a_2) = \{c_1, c_2, c_3\}$.
Outputs	**ACU-level verdicts:**	$s_1 = 1$ (supported), $t_1 = \varnothing$; $s_2 = 0$ (unsupported), $t_2 =$ **EntE**.
Aggregation		$\text{FAIR\_Score} = \frac{1}{\lvert U \rvert} \sum_{k=1}^{\lvert U \rvert} s_k = \frac{1}{2}(1+0) = 0.5$.

following prior work, we adopt the FTSOTA split, which focuses on summaries produced by strong fine-tuned models and is therefore more challenging for conventional factuality metrics. LLMSumEval [1] contains human judgments on summaries generated by large language models under zero-shot and few-shot settings, representing a particularly difficult regime where fluent hallucinations are prevalent.

Evaluation protocol. Following SummaC [17], we standardize each benchmark into a binary factual consistency classification task by selecting a decision threshold on the validation set and reporting balanced accuracy (BACC) on the test set to mitigate label imbalance. In addition to classification performance, we compute Pearson r and Spearman ρ correlations between metric scores and human judgments to evaluate how well each automatic metric reflects human-perceived faithfulness.

B. Implementation Details

We instantiate each module in FAIR with a dedicated backbone aligned with its role in the pipeline. For logic-aware segmentation, we employ a Llama2-70B [26] model fine-tuned on ORCA [27] to better capture discourse-level relations and support boundary decisions that respect logical structure. For semantic-aware segmentation and evidence ranking, we use bge-reranker-v2-m3 to compute semantic similarity. We set the similarity threshold to 0.78, the chunk size to 500 tokens, and the overlap to 50 tokens to balance contextual completeness and retrieval precision. For verification, we adopt Qwen2.5-7B and prompt it to output (1) a binary factuality label for each ACU and (2) an explicit error type from the taxonomy in Table I when the ACU is unsupported.We prompt the verifier to output an ACU-level binary label and an explicit error type for unsupported units. The prompt template is provided in Fig. 3.

We optimize trainable components with AdamW for 10 epochs (batch size 16, learning rate 2×10^{-5}, dropout 0.1), apply gradient clipping at 1.0 for stability, and use early stopping based on development performance to prevent overfitting.

C. Baseline Methods

We compare FAIR with representative factuality metrics spanning different paradigms. DAE [18] represents dependency-based entailment evaluation, which checks factuality through structured entailment signals. QuestEval [20] and QAFactEval [19] are QA-based metrics that probe factual consistency by asking questions derived from the summary and validating answers against the source. We also include LLM-based evaluators, including ChatGPT-ZS (zero-shot prompting) and more structured prompting approaches, such as G-Eval [28] and BelugaEval [28], which have shown strong performance but may be sensitive to prompts and evidence selection. This suite provides a comprehensive comparison against both traditional and recent LLM-driven factuality evaluation strategies.

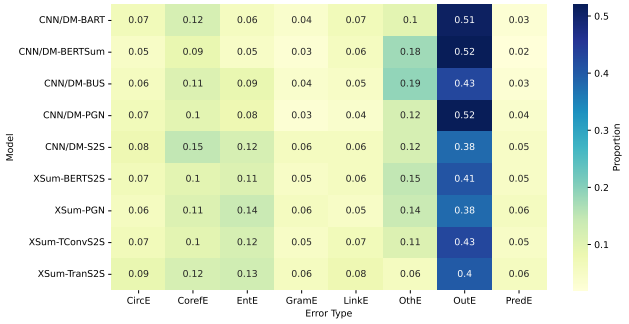

Fig. 4. Proportion of summaries containing factual errors, with a breakdown by error categories. Detailed definitions of error categories are provided in Table I.

D. Comparison Results

We adopt balanced accuracy (BACC) as the primary metric for binary factuality classification. As shown in Table III, QA-based evaluators are already competitive; for example, QAFactEval achieves 83.0 ± 1.7 BACC on SummEval and performs comparably on AggreFact-FTSOTA. However, these approaches rely on fixed question generation and answer matching, which can be sensitive to how multiple claims are packaged within a sentence and thus provide limited error localization. A representative LLM-based evaluator, G-Eval, also yields strong results (e.g., 81.9 ± 1.5 on SummEval), but it primarily produces scalar judgments without structured unit-level diagnostics. In contrast, FAIR achieves the best performance across all evaluated settings (e.g., 86.5 ± 2.4 on SummEval and 88.5 ± 1.3 on LLMSumEval), with 95% confidence intervals indicating statistically stable improvements. These results support the effectiveness of jointly controlling claim granularity and evidence selection for robust factuality evaluation.

TABLE III
BALANCED ACCURACY ON SUMMARIZATION BENCHMARKS WITH 95% CONFIDENCE INTERVALS.

	SUMMEVAL	AGGREFACT-FTSOTA		LLMSUMMEVAL	
		CNN/DM	XSum	CNN/DM	XSum
DAE	64.8 ± 2.4	65.4 ± 4.4	70.2 ± 2.3	84.6 ± 1.7	72.9 ± 1.6
QuestEval	73.8 ± 2.6	70.2 ± 3.2	59.5 ± 2.7	86.5 ± 1.9	75.1 ± 1.5
QAFactEval	83.0 ± 1.7	67.8 ± 4.1	63.9 ± 2.4	68.3 ± 3.4	62.3 ± 2.0
ChatGPT-ZS	-	56.3 ± 2.9	62.7 ± 1.7	-	-
G-Eval	81.9 ± 1.5	-	-	-	-
BelugaEval	81.1 ± 1.6	56.1 ± 2.8	66.1 ± 1.7	77.0 ± 2.0	62.8 ± 1.7
Ours	**86.5 ± 2.4**	**70.4 ± 1.3**	**69.5 ± 1.9**	**88.5 ± 1.3**	**79.4 ± 2.5**

TABLE IV
CORRELATION OF AUTOMATIC METRICS WITH HUMAN JUDGMENTS ON SUMMEVAL, AGGREFACT-FTSOTA, AND LLMSUMEVAL. WE REPORT PEARSON r AND SPEARMAN ρ. HIGHER VALUES INDICATE STRONGER ALIGNMENT WITH HUMAN EVALUATION.

Metric	SummEval		AggreFact-FTSOTA		LLMSumEval	
	r	ρ	r	ρ	r	ρ
ROUGE-1	0.220	0.237	–	–	–	–
ROUGE-2	0.174	0.189	0.459	0.418	0.097	0.083
ROUGE-L	0.221	0.224	0.357	0.324	0.024	–0.011
BERTScore	0.304	0.203	0.576	0.505	0.024	0.008
BARTScore	0.290	0.311	0.735	0.680	0.184	0.159
MoverScore	0.319	0.166	0.414	0.347	0.054	0.044
FactCC	–	–	0.416	0.484	0.297	0.259
QAGS	–	–	0.545	–	0.175	–
QuestEval	0.392	0.420	–	–	–	–
UniEval	–	–	0.682	0.662	0.461	0.488
G-Eval-3.5	0.385	0.386	0.477	0.516	0.211	0.406
Ours	**0.468**	**0.441**	**0.599**	**0.611**	**0.472**	**0.460**

Alignment with human judgments. Table IV reports Pearson r and Spearman ρ correlations with human annotations. Traditional similarity metrics exhibit limited agreement with humans (e.g., ROUGE-1 achieves only $r = 0.220$ on SummEval), motivating the use of factuality-oriented evaluators. Recent LLM-based metrics provide substantially stronger correlations on some benchmarks (e.g., UniEval on AggreFact-FTSOTA), but they typically return aggregated scores without unit-level error attribution. In contrast, FAIR achieves consistently high alignment across all three datasets (e.g., $r = 0.468,\ \rho = 0.441$ on SummEval), outperforming similarity-based and classifier-based baselines and remaining competitive with the strongest LLM-based evaluators (Table IV). Beyond correlation, FAIR outputs ACU-level support decisions and explicit error categories, enabling an interpretable diagnosis of where and why factual inconsistencies arise rather than collapsing evaluation into a single scalar.

E. Ablation Study

To investigate the contribution of different components in FAIR, we perform ablation experiments by removing the logic-aware segmentation (*w/o Logic*) and semantic-aware segmentation (*w/o Semantic*). The results are summarized in Fig. 5 .

Across all benchmarks (SummEval, AggreFact, and LLM-SumEval), removing either module leads to a clear performance drop. In particular, discarding the logic-aware module results in degraded accuracy on XSum, indicating that granularity-controlled atomic decomposition is essential for detecting factual errors. Similarly, removing semantic-aware segmentation substantially decreases performance on CNN/DM, suggesting that coherent document chunking is critical for accurate evidence retrieval.

These findings demonstrate the complementary roles of logic-aware and semantic-aware segmentation: the former ensures the precise decomposition of summary facts, while the latter maintains the semantic integrity of document evidence. These results confirm that the joint design of logic- and semantic-aware segmentation is complementary and indispensable for achieving FAIR's superior robustness and generalization across benchmarks.

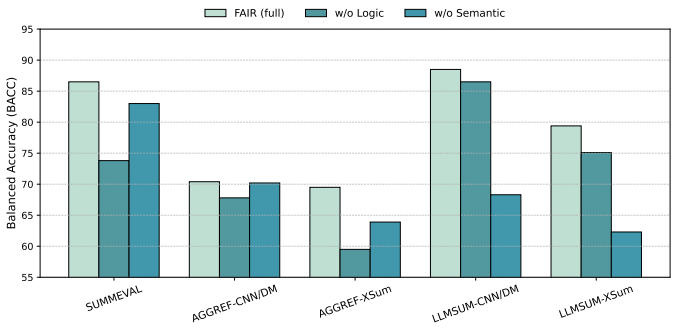

Fig. 5. Ablation Results on Five Datasets

V. CONCLUSION

In this work, we revisit factual consistency evaluation for summarization from the perspective of granularity control. We identify granularity mismatch as a hidden yet systematic source of evaluation bias: under-segmentation can merge logically independent facts and mask localized errors, whereas over-segmentation can fragment semantically interdependent statements and break relational semantics. This observation suggests that reliable and interpretable evaluation requires explicit granularity-controlling modeling rather than a fixed granularity assumption. The proposed FAIR jointly performs logic-aware segmentation and semantic-aware segmentation, enabling retrieval-enhanced ACU verification with explicit error typing. Extensive experiments on three benchmarks show that FAIR improves alignment with human judgments, reduces evaluation bias, and achieves competitive or state-of-the-art performance baselines. This work may bring new insights to other generation tasks beyond summarization (e.g., long-form QA and report generation) where granularity-induced hidden bias in evaluation may similarly arise.

VI. ACKNOWLEDGEMENTS

This research is supported by the National Key R&D Program of China (No. 2023YFC3303800).

References

[1] Y. Liu et al., "Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization," *ArXiv*, 2023.

[2] Z. Luo, Q. Xie, and S. Ananiadou, *Chatgpt as a factual inconsistency evaluator for text summarization*, 2023. arXiv: 2303.15621 [cs.CL].

[3] M. Peyrard, "Studying summarization evaluation metrics in the appropriate scoring range," in *ACL*, 2019, pp. 5093–5100.

[4] S. Min et al., "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation," *ArXiv*, 2023.

[5] A. Scirè, K. Ghonim, and R. Navigli, "Fenice: Factuality evaluation of summarization based on natural language inference and claim extraction," *ArXiv*, 2024.

[6] D. Wan, K. Sinha, S. Iyer, A. Celikyilmaz, M. Bansal, and R. Pasunuru, "Acueval: Fine-grained hallucination evaluation and correction for abstractive summarization," in *ACL*, 2024, pp. 10 036–10 056.

[7] Y. Liu et al., "Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation," *ArXiv*, 2022.

[8] R. Qu, R. Tu, and F. S. Bao, "Is semantic chunking worth the computational cost?" In *Findings of the Association for Computational Linguistics: NAACL 2025*, Apr. 2025, pp. 2155–2177, ISBN: 979-8-89176-195-7.

[9] J. Liu, Z. Shi, and A. Lipani, "Summequal: Summarization evaluation via question answering using large language models," in *ACL*, 2024, pp. 46–55.

[10] D. Lei et al., "Chain of natural language inference for reducing large language model ungrounded hallucinations," *ArXiv*, 2023.

[11] L. Tang et al., "Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors," *ArXiv*, 2022.

[12] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, "Summeval: Re-evaluating summarization evaluation," *TACL*, vol. 9, pp. 391–409, 2021.

[13] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[14] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *ArXiv*, 2019.

[15] W. Zhao, M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, and S. Eger, "Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance," *ArXiv*, 2019.

[16] W. Kryscinski, B. McCann, C. Xiong, and R. Socher, "Evaluating the factual consistency of abstractive text summarization," Nov. 2020, pp. 9332–9346.

[17] P. Laban, T. Schnabel, P. N. Bennett, and M. A. Hearst, "Summac: Re-visiting nli-based models for inconsistency detection in summarization," *TACL*, vol. 10, pp. 163–177, 2022.

[18] T. Goyal and G. Durrett, "Evaluating factuality in generation with dependency-level entailment," *ArXiv*, 2020.

[19] A. R. Fabbri, C.-S. Wu, W. Liu, and C. Xiong, "Qafacteval: Improved qa-based factual consistency evaluation for summarization," *ArXiv*, 2021.

[20] T. Scialom et al., "Questeval: Summarization asks for fact-based evaluation," *ArXiv*, 2021.

[21] T. Goyal and G. Durrett, "Annotating and modeling fine-grained factuality in summarization," *ArXiv*, 2021.

[22] J. Yang, H. Liu, W. Guo, Z. Rao, Y. Xu, and D. Niu, "Reassess summary factual inconsistency detection with large language model," in *KnowLLM*, 2024, pp. 27–31.

[23] J. Yang, S. Yoon, B. Kim, and H. Lee, "Fizz: Factual inconsistency detection by zoom-in summary and zoom-out document," *ArXiv*, 2024.

[24] M. Wanner, S. Ebner, Z. Jiang, M. Dredze, and B. Van Durme, "A closer look at claim decomposition," Mexico City, Mexico: ACL, Jun. 2024, pp. 153–175.

[25] X. Li, C. Zhu, L. Li, Z. Yin, T. Sun, and X. Qiu, "LLatrieval: LLM-verified retrieval for verifiable generation," Jun. 2024, pp. 5453–5471.

[26] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," *ArXiv*, 2023.

[27] S. Mukherjee, A. Mitra, G. Jawahar, S. Agarwal, H. Palangi, and A. Awadallah, "Orca: Progressive learning from complex explanation traces of gpt-4," *ArXiv*, 2023.

[28] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-eval: Nlg evaluation using gpt-4 with better human alignment," *ArXiv*, 2023.