Revisiting Faithfulness Annotations for Long-form Summaries

Yang Zhong[1], Yang Janet Liu[*2], Diane Litman[*1,3]
[1]Department of Computer Science, [2]Department of Linguistics
[3]Learning Research & Development Center
University of Pittsburgh, Pittsburgh, PA, USA
{yaz118,jal787,dlitman}@pitt.edu

Abstract

Benchmarks for long-form summaries (four or more sentences) generated by language models increasingly serve as gold-standard references for developing, evaluating, and comparing faithfulness-checking systems. As their influence grows, understanding the challenges of annotating faithfulness errors within long, discourse-rich summaries becomes critical. We revisit three benchmarks spanning diverse text types and contrasting annotation designs. Using a discourse-aware evaluation framework together with human auditing, we identify cases where benchmark labels may be unreliable. Manual verification shows that 3.4%-5.4% of sentence-level labels warrant revision due to discourse-level inconsistencies that standard annotation procedures overlook. We introduce a taxonomy of five recurring annotation error types, propose revised labels, and show that correcting these cases leads to meaningful shifts in system rankings. We conclude with recommendations for future annotation practices.

1 Introduction

Faithfulness benchmarks for AI-generated summaries have become de facto gold-standard references for developing and comparing automatic fact-checking systems (Kryscinski et al., 2019; Fabbri et al., 2021; Tang et al., 2023), making their annotation reliability a foundational concern. Evaluating long-form summaries presents distinct challenges. Unlike news-domain benchmarks, where summaries typically contain one to three sentences, long-form summaries often exceed 100 words and synthesize information from multiple parts of the source document. This increased complexity requires annotators to assess not only the accuracy of individual facts but also the coherence and validity of relationships among them. Prior work has shown that such demands impose substantial cognitive load and can reduce inter-annotator agreement (Krishna et al., 2023). Moreover, faithfulness errors in long-form summaries increasingly arise from subtle distortions in meaning or framing (Zhong and Litman, 2025b), rather than from surface-level inconsistencies (Goyal et al., 2022).

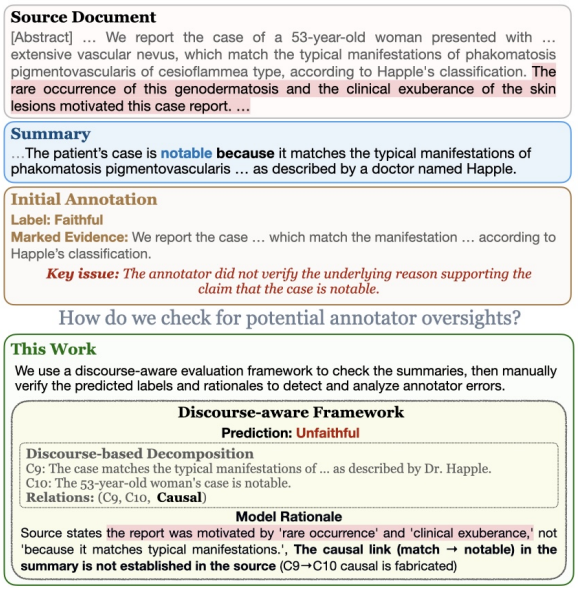

Figure 1: **Human annotators overlooked discourse-level errors.** The source document attributes the case report's motivation to "the rare occurrence" and the "clinical exuberance". However, the annotator (a domain expert) relied on a preceding evidence sentence and failed to evaluate the *causal relationship* underlying the case's significance.

A closer examination of long-form summaries reveals that many problematic cases stem not from incorrect individual facts, but from whether the *relationships* between them are sufficiently established. For example, Figure 1 shows that a human annotator labeled a summary sentence as faithful because the individual facts appear in the source. However, the annotation overlooks whether the causal relationship—namely, that "the matching of the description of a patient's case" makes it "notable"—is actually supported by the document. Such errors

[*]Equal senior contribution.

Proceedings of the 20th Linguistic Annotation Workshop (LAW XX), pages 158–172

are difficult to detect when annotation protocols prioritize verifying the presence of discrete facts over assessing how those facts are connected. Similar issues arise when summaries alter temporal ordering, omit hedging that overstate certainty, or recast factual content with interpretive language.

Motivated by these challenges, we revisit three recent benchmarks for evaluating faithfulness in long-form summaries, which span diverse domains and annotation designs. Using a discourse-aware evaluation framework together with complementary LLM-based analysis methods, we conduct a systematic audit of sentences where automatic evaluators disagree with the gold annotations. For these cases, we perform manual verification to assess whether the original labels remain justified (§4).

Our reassessment reveals that 3.4–5.4% of sentence-level annotations require revision upon closer examination (§5). To characterize these inconsistencies, we propose a taxonomy of five recurring annotation error categories grounded in discourse analysis: *fabricated relational links*, *scope and attachment errors*, *temporal and state change errors*, *hedging removal*, and *semantic reframing* (§6). We then analyze the downstream impact of these revisions on system evaluation (§7), finding that even small corrections yield clearer separation between competing evaluation approaches. This suggests that annotation inconsistencies can obscure meaningful differences in system performance.[1] Finally, we offer recommendations for future annotation efforts in long-form summarization, including the adoption of discourse-aware tools for quality assurance and improved design of error taxonomies and annotation frameworks (§8).

2 Related Work

Annotation Reliability in Faithfulness Benchmarks. Recent studies have raised concerns about the reliability of "gold-standard" annotations used in faithfulness evaluation benchmarks. Laban et al. (2023) conduct a manual analysis of model–dataset disagreements and estimate that at least 6% of samples in AggreFact (Tang et al., 2023), a widely used benchmark, are mislabeled. Similarly, Seo et al. (2025) report that in aggregated factuality datasets, 9.1% of examples are ambiguous and 6.6% appear mislabeled. Beyond annotation noise, Godbole and Jia (2025) show that state-of-the-art LLM-based faithfulness evaluators often rely on surface-level similarity, failing to properly verify claims that require aggregating information across distant parts of a source document. This limitation calls into question the reliability of methods that rely on LLM-generated rationales to guide human assessments, such as Lee et al. (2024). Taken together, these findings highlight the need for a more careful and systematic examination of annotation reliability in faithfulness benchmarks. *We address this gap through a discourse-aware reassessment of recent long-form summary datasets, supported by targeted human verification.*

Long-form Summary Annotation Granularity. As benchmarks expand to longer summaries, researchers have increasingly adopted fine-grained annotation schemes, including sentence-level error taxonomies (Koh et al., 2022) and span-level annotations with taxonomies designed to capture coherence-related errors in narrative texts (Goyal et al., 2022). Krishna et al. (2023) further show fine-grained clause-level annotations yield better inter-annotator agreement than holistic summary-level judgment. More recent long-form summary benchmarks have continued this trend by adopting sentence-level annotations, including both binary (Subbiah et al., 2024; Fang et al., 2024) and fine-grained labels (Ding et al., 2025). Yet even fine-grained taxonomies operate primarily at the sentence level and rarely require annotators to verify relations between claims/sentences (e.g., causal or temporal). *Our work investigates this overlooked limitation by examining how human annotations fail to capture discourse-level inconsistencies.*

Discourse Analysis in Faithfulness Evaluation. Pagnoni et al. (2021) incorporate discourse link errors into their FRANK typology, marking an early effort to account for discourse-level phenomena in faithfulness evaluation. Building on this perspective, Zhong and Litman (2025a) show that faithfulness errors in long-form summaries correlate with specific discourse features such as nuclearity. Subsequent work has further incorporated discourse relations into structured faithfulness evaluation (Zhang et al., 2025). *In contrast, our work leverages a discourse-informed framework to produce fine-grained labels with interpretable, discourse-grounded rationales. We employ this framework as an auditing tool to surface systematic annotation errors that are often overlooked by standard evaluation procedures.*

[1] The revised datasets and codebase are available at https://github.com/cs329yangzhong/faithfulEval4LongSumm.

	Text Diversity	Human Annotation				Dataset Statistics		
	Doc.Src	Taxonomy	Annotator	Protocol	Aggregation	# Summ	# Sent	Sum.Word
StorySumm	narratives	binary	6 crowd workers + 3 experts	hybrid human-AI	unanimity + adjudication	96	580	139
VeriGray	news	7-category	2 grad. students	double annot. + LLM review	disagreement review	412	2,044	122
FaReBio	scientific	binary	2 medical doctors	single annotation	n/a	175	1,445	198

Table 1: **Overview of the three analyzed benchmarks.** We report the source document type (Doc.Src), annotation taxonomy, annotator type, protocol, and label aggregation strategy, along with key statistics: number of summaries (# Summ), number of annotated sentences (# Sent), and average summary length in words (Sum.Word).

3 Datasets

We analyze three faithfulness benchmarks for LLM-generated summaries which vary in domain, annotation protocol, and label taxonomy: StorySumm, VeriGray, and FaReBio. Covering narrative, news, and scientific texts and applying distinct annotation strategies, they allow us to examine how annotation design affects labeling reliability. Table 1 summarizes their key characteristics.

StorySumm (Subbiah et al., 2024) contains 96 narrative summaries (580 annotated sentences). Each summary sentence is first labeled by three[2] crowd workers with binary faithfulness judgments, with explanations provided for unfaithful cases. The annotation interface includes an "N/A, commentary" option for sentences that interpret narrative themes rather than summarize story plots. To enhance label quality, the dataset adds (i) an *expert* setting with author adjudication at the summary level, and (ii) a *hybrid* setting where GPT-4 generates potential inconsistencies to guide a second round of crowd annotations. Final labels are derived via unanimity and adjudication, yielding a public version where some sentence labels diverge from the original unanimity votes. For our analysis, we use the raw per-annotator files, including individual votes and explanations.

VeriGray (Ding et al., 2025) contains 412 news article summaries (2,044), annotated by graduate students following a fine-grained taxonomy (Appendix A.1). The annotation protocol involves double annotation, followed by LLM-based reviews with prompts adopted from Seo et al. (2025), and a final annotator discussion to resolve disagreements. Here we analyze the initial dataset release (October 2025).

FaReBio (Fang et al., 2024) contains 175 summaries (1,445 sentences) of biomedical research abstracts and introduction texts, annotated by two medical doctors for faithfulness and factual hallucination. Annotations include highlighting evidence spans and free-text explanations. Only a small subset of the data is doubly annotated (34 sentences, Cohen's $\kappa = 0.48$); the remaining sentences are singly annotated. Further details on annotation procedures and dataset construction of the three datasets are provided in Appendix A.

4 Reassessment Procedure

Our central hypothesis is that discourse-level inconsistencies (fabricated relations, temporal reordering) are overlooked by standard annotation procedures, which prioritize verifying individual facts over the relationships between them. To test this, we design a three-stage reassessment pipeline grounded in discourse analysis.

Auditing Tool. Our primary auditing tool is DeScribe,[3] a discourse-aware faithfulness evaluation framework grounded in Rhetorical Structure Theory (RST, Mann and Thompson 1988). Given a source document and a summary, DeScribe decomposes the summary into atomic claims while preserving discourse relations (e.g., causal, attribution, temporal, contrast), and performs structured inference to verify both claims *and* their intra- and inter-sentential relations against the source. This process yields fine-grained error labels with explicit rationales, identifying cases where facts are correct but the links between them are unsupported. Figure 2 (A) shows this for the Figure 1 example.

Stage 1: Automated Flagging. We apply DeScribe to evaluate all summary sentences across the three datasets. A sentence is flagged whenever DeScribe's prediction disagrees with the gold label. We additionally apply two complementary

[2]This accounts for the annotator label construction in the original paper (Subbiah et al., 2024). The authors include two additional annotation settings (Stage 2 in §4), with a total of six crowd workers and three experts involved in the entire annotation process (Table 1).

[3]https://anonymous.4open.science/r/Describe-2C68/

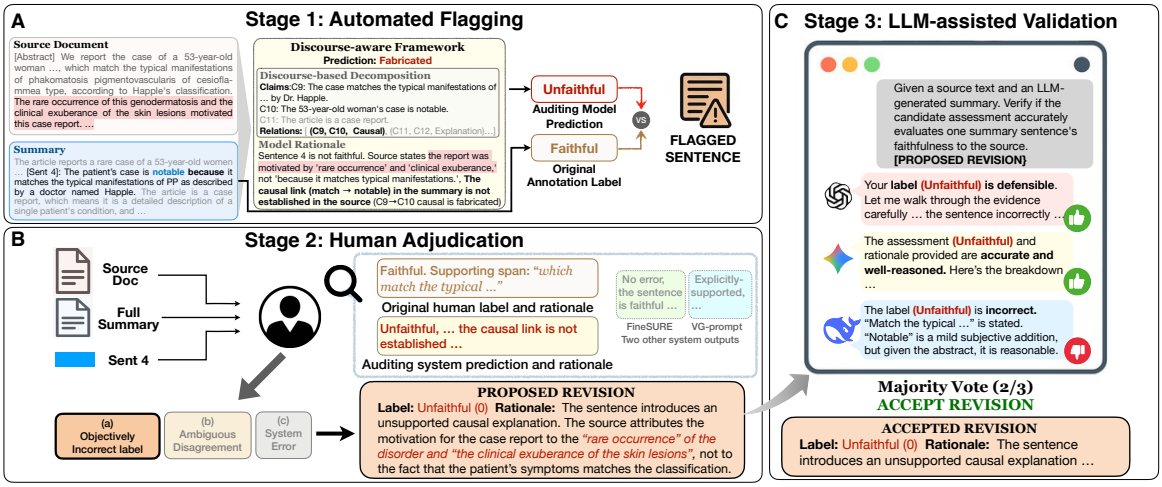

Figure 2: **Three-stage reassessment pipeline on a biomedical summary from FAREBIO.** Panel **(A)**: a discourse-aware framework flags a fabricated causal link absent from the source. Panel **(B)**: human adjudication reviews the source evidence and all system rationales and proposes a revised label with corresponding rationales. Panel **(C)**: Three independent LLMs verify the proposed revision by majority vote. The example highlights a recurring pattern: individual facts are correct, but the relation connecting them is unsupported.

baselines to the flagged sentences to provide diverse rationales for Stage 2.

Stage 2: Human Adjudication. For each flagged instance, the first author reviewed the source document, full summary, gold label, annotator explanations (if available), and the auditing system's predictions and rationales. To facilitate manual validation, we also apply two LLM-based faithfulness checkers with different prompting strategies—FineSURE (Song et al., 2024) and VG-prompt (Ding et al., 2025)—on those flagged sentences.[4] All systems use GPT-5 (medium reasoning effort) as the backbone. Each case is classified as: (a) *objectively incorrect*—the gold label contradicts guidelines or textual evidence; (b) *ambiguous disagreement*—both annotation and system rationale are defensible under reasonable interpretation, and (c) *system error*—the gold label is correct and the faithfulness checking systems err. For objectively incorrect cases, we propose revised labels aligned with the dataset taxonomy, with free-text rationales grounded in textual evidence, informed by DESCRIBE's discourse-based rationales (see Figure 2 (B) for a worked example).

Stage 3: LLM-assisted Verification. To mitigate potential adjudication bias from a single annotator, all proposed revisions for "objectively incorrect" cases and their rationales are independently verified by three LLMs through web-based chats: `Gemini-3-Pro`, `GPT-5.2-Thinking`, and `DeepSeek-V3.2-Thinking`.

As illustrated in Figure 2 (C), a revision is accepted only if it is supported by verifiable textual evidence (e.g., a misattribution, a numeric discrepancy, or an unsupported discourse relation) and corroborated by at least two models. This process ensures accepted changes reflect clear annotation inconsistencies rather than borderline disagreements.

5 Reassessment Outcomes

Table 2 presents detailed auditing results by category after applying our three-stage reassessment pipeline. We leave the ambiguous cases unchanged, focusing on clearly and objectively incorrect labels. We propose revised labels for 28 sentences in STORYSUMM (4.8% of 580), 69 in VERIGRAY (3.4% of 2,044), and 78 in FAREBIO (5.4% of 1,445). In STORYSUMM, where the raw per-annotator files are available, 15 of the 28 revised cases were flagged by at least one original crowd worker, indicating that valid minority signals were present but suppressed during label aggregation.

Dataset-specific Patterns. Revision patterns are not random but align with dataset-specific annotation protocols. In STORYSUMM, 21 of 28 revisions involve sentences originally labeled as faithful, reflecting the unanimity rule that labels a sentence as unfaithful only with full agreement—biasing the dataset toward under-detection of subtle errors. In VERIGRAY, most revisions occur at the

[4] Appendix B includes prompting details.

Dataset	Total Sent.	Flagged	Obj. Incorrect	Ambiguous	Inconsistency Rate
StorySumm	580	91	28	9	4.8%
VeriGray	2,044	98	69	4	3.4%
FAREBIO	1,445	115	78	5	5.4%

Table 2: Full auditing results after applying our three-stage reassessment pipeline presented in §4.

boundary between *Generally-Supported* and *Fabricated*, a distinction requiring judgment of when semantic strengthening becomes unsupported. In FAREBIO, where most sentences are annotated individually by medical doctors, revisions appear to reflect annotator-specific tendencies rather than aggregation effects. We observe repeated cases where surface-plausible comparative claims are accepted without verifying whether hedging in the source text has been preserved, suggesting that domain expertise may encourage acceptance of clinically consistent interpretations.

Prevalence of Human Disagreement in StorySumm. Using the official annotation files shared by the authors of StorySumm, we analyze the crowd-sourced annotation process and observe substantial disagreement among annotators. Specifically, 151 (out of 580) sentences received at least one label of "No" during annotation. Among these cases, the distribution of unfaithfulness votes—corresponding to cases where one, two, or all three annotators labeled a sentence as "No"—is 66.2%, 20.5%, and 13.2%, respectively. This suggests that unanimous agreement on unfaithfulness is relatively uncommon.

We further examine how these annotator votes are reflected in the aggregated sentence-level labels. For sentences where only one of three annotators assigned a "No" label, only 19% are ultimately labeled as "No" in the final dataset. In contrast, when two of three annotators labeled the sentence as "No", the final label aligns with the majority judgment in 84% of cases. These findings indicate that the aggregation procedure substantially suppresses minority disagreement signals, leaving a large proportion of partially disagreed cases labeled as faithful. This observation suggests the presence of non-trivial annotation ambiguity and motivates targeted validation of these disputed instances.

6 Error Analysis

To characterize recurring patterns among the revised cases, we develop a taxonomy of five error types through qualitative analysis. This analysis identifies systematic sources of faithfulness errors that persist despite surface-level factual overlap. The resulting categories draw on established concepts from discourse processing and prior error taxonomies, including relational entailment, scope and attachment resolution, and temporal ordering (Mann and Thompson, 1988; Pagnoni et al., 2021).[5] Table 3 presents the five categories with diagnostic cues, discourse-based interpretations, and illustrative examples. Figure 3 shows their normalized distribution across datasets.[6] We highlight and discuss major observations across datasets below.

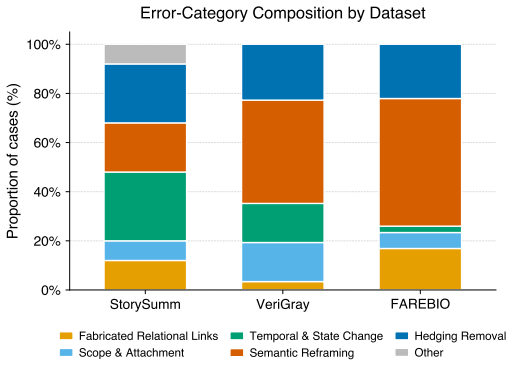

Figure 3: Normalized distribution of annotation error types across benchmarks.

Semantic reframing and hedging removal dominate across datasets. In VeriGray and FAREBIO, most revisions involve either *semantic reframing*, which replaces factual descriptions with interpretive or evaluative language, or *hedging removal*, which omits epistemic qualifiers. Together, these categories account for roughly 70% of revisions. Because both preserve surface plausibility, they are difficult to detect when annotators judge whether content "sounds right" rather than whether the source framing is faithfully preserved.

We identify three subtypes of semantic reframing: (1) *interpretive labeling* (e.g., adding evalua-

[5] We used LLM-assisted summaries of the revised cases to surface candidate patterns, which were then manually consolidated and refined into the final taxonomy.

[6] Two additional StorySumm cases involve named entity errors outside the taxonomy and were corrected by the dataset authors after our notification.

Category	Typical cues	Definitions	Examples
Fabricated Relational Links	*because, due to, leading to, resulting in; but, instead of; if/unless*	Errors often involve **unsupported relations** (causal, contrastive, conditional) rather than missing facts.	`Summary:` "Casey leaves with the money **but** feels confused about the situation." `Evidence:` Casey felt "confused" when asked to sign a name that wasn't hers on her painting. After the argument she signed her own name, causing the requester to become "distraught" and "set about destroying the painting." Casey then "said she was sorry and asked to be paid," and "took the money and left." `Error:` The word "but" fabricates a contrastive relation between the money and the confusion; the source presents them as separate, sequential events with no opposition.
Scope & Attachment	Long multi-clause sentences; distant modifiers	Errors arise when connectives or modifiers attach to the **wrong proposition**, yielding false support under span matching.	`Summary:` Judge Ian Pearson remanded her into custody, stating she would likely be held at a male prison **due to** the risk of self-harm and non-attendance. `Evidence:` The judge said the defendant "would be a risk to herself and a risk of failing to attend," thus ordering "remand in custody." The reasoning for a male prison was that the defendant "had not made any physical changes or enhancements to her body or taken any medication." `Error:` The summary incorrectly attaches the justification for the remand decision (self-harm risk, non-attendance) to the male-prison decision; each rationale belongs to a different proposition in the source.
Temporal & State Change	*then/after/before*; verbs like *realize/discover/decide*	Errors occur when the correct fact is placed at the **wrong temporal position** in the narrative.	`Summary:` [Context:] When he hints at the gift, she gives him a blank stare and denies having ever done so. [Target:] They realize her memory is rapidly deteriorating. `Evidence:` The narrator recounted, "A few months ago I started to notice the memory problems"; the family shared the realization "we were so saddened about the Great Forgetting." `Error:` The source places the narrator's judgment about rapid deterioration *before* the hint exchange, not as a new realization triggered by it. The summary reverses the causal–temporal order.
Hedging Removal	*may aid* vs *are needed*; *mostly* vs *all*; *seemed* vs *is*	Errors arise when hedges or attributions are dropped, inflating **certainty beyond the source**.	`Summary:` The degradation correction **is** a manual process based on assumed knowledge of the sensor hardware. `Evidence:` "Degradation correction has been **mostly** a manual process based on assumed knowledge of the sensor hardware." `Error:` The summary drops the hedge "mostly" and states the claim categorically, inflating certainty beyond what the source supports.
Semantic Reframing	interpretive wording (e.g., *high-risk, effective, promising*); terminology reinterpretation; paraphrase drift	Errors arise when the summary **reframes a factual statement with an interpretive or evaluative label** not explicitly supported by the source.	`Summary:` It could also correctly identify artificially mutated viral genomes as **high-risk**. `Evidence:` "The artificial negative data with the replacement of the coding region of the spike protein were also predicted correctly (100% accuracy)" `Error:` By replacing the spike protein (the key to cross-species infection), the researchers created genomes that should not cause a pandemic. The model correctly identified these as negative (low risk),

Table 3: **Five recurring categories of overlooked annotation inconsistencies across datasets.** We present diagnostic cues, discourse-based detection insights, and illustrative examples with supporting source evidence.

tive terms not stated in the source like "high-risk"), (2) *domain-knowledge equivalence* (substituting terms based on unstated background knowledge), and (3) *interpretive inference* (drawing broader conclusions from observations). In FAREBIO, domain expertise appears to amplify this issue, as annotators naturally accept clinically plausible interpretations without explicit textual support. Similarly, removing hedges (e.g., "mostly," "felt like") shifts qualified claims into a definitive claim—an epistemic change that is easily overlooked despite preserved core content.

Backbone	Method	StorySumm				VeriGray				FaReBio			
		Orig.	Rank	Rev.	Rank	Orig.	Rank	Rev.	Rank	Orig.	Rank	Rev.	Rank
Gemini-3-Flash	Fewshot	64.9	3	66.8	3	72.7	3	72.6	3	**70.9**	1	73.6	2▽
	FineSURE	**70.1**	1	**69.8**	1	72.2	4	71.8	4	69.8	3	**74.2**	1△
	VG-prompt	61.8	4	63.1	4	73.2	2	73.9	2	70.0	2	71.5	3▽
	Describe	65.0	2	67.1	2	**75.5**	1	**76.8**	1	68.2	4	70.2	4
GPT-5	Fewshot	72.0	3	75.8	3	80.9	3	81.3	4▽	**79.1**	1	85.5	3▽
	FineSURE	**75.0**	1	77.2	2▽	81.0	2	81.6	3▽	77.2	3	82.7	4▽
	VG-prompt	72.9	2	75.5	4▽	80.6	4	82.5	2△	78.5	2	85.7	2
	Describe	71.4	4	**77.5**	1△	**81.5**	1	**84.9**	1	76.0	4	**87.3**	1△

Table 4: **Balanced accuracy (%) and system rankings before (Orig.) vs. after (Rev.) label revision across three datasets and two backbone models.** Best results per model group are **bold**. △ and ▽ mark methods whose rank improved or dropped after label revision. Predictions are unchanged; differences reflect corrected annotations. These shifts illustrate how annotation inconsistencies can obscure meaningful system rankings.

StorySumm exhibits a distinct narrative-driven error profile. Unlike the other datasets, revisions in StorySumm are more evenly distributed across categories, with *temporal and state-change* errors being most prevalent. This reflects the narrative text type: stories rely heavily on event order and the timing of character realizations, so misplacing an otherwise correct fact at the wrong narrative moment can distort its significance. This observation aligns with prior work emphasizing the importance of temporal discourse structure in narrative understanding (Genette, 1983; Hamilton et al., 2025). In contrast, temporal errors are rare in VeriGray and FaReBio, where news and scientific texts are organized more thematically than chronologically, reducing opportunities for reordering errors. *Fabricated relational links* also occur relatively frequently, particularly in narrative summaries that compress complex plot structure into simplified causal or contrastive structures not present in the source document (e.g., the inserted "but" in the Casey example, Table 3, row 1).

A common thread: correct facts, unsupported relations. Across all categories and datasets, revised cases share a consistent pattern: individually correct facts paired with unsupported or distorted *relations* between them. For *fabricated relational links*, both cause and effect may appear in the source, but the causal connection is invented. Similarly, the introduction of contrastive relations (first example in Table 3), using discourse markers such as *instead of* or *but*, creates unsupported contrasts between otherwise valid facts. For *scope and attachment* errors, all facts are present, but modifiers attach to the wrong proposition. This pattern helps explain why such inconsistencies evade standard annotation: verifying that individual facts appear in the source does not guarantee that the relationships between them, such as causal, temporal, attributive, or evaluative, are also supported.

Binary taxonomies (faithful/unfaithful) exhibit higher revision rates (4.8–5.4%) than VeriGray's finer-grained seven-category scheme (3.4%), suggesting that additional distinctions help annotators recognize borderline cases. However, even fine-grained taxonomies remain vulnerable when applied at the sentence level without explicit verification of discourse relations. These observations motivate our recommendations in §8.

7 Annotation Revisions Enhance System Comparisons

Experimental Setup. We evaluate downstream effects by comparing system performance under original vs. revised labels for four faithfulness checking methods: Fewshot (Seo et al., 2025), FineSURE (Song et al., 2024), VG-prompt (Ding et al., 2025), and Describe using two representative backbone LLMs (`Gemini-3-Flash` and `GPT-5`). Performance is measured by sentence-level balanced accuracy, and we report system rankings per backbone and dataset. Experimental details are documented in Appendix C.

Observations. Table 4 shows that correcting a small number of annotations leads to meaningful shifts in system rankings. Of 24 system–dataset–backbone slots, 12 exhibit rank shifts after revision, with the effect concentrated under `GPT-5` (9 of 12 slots) compared to `Gemini-3-Flash` (3 of 12). Under the original labels, system rankings are inconsistent: FineSURE (`with GPT-5`) ranks first on StorySumm, Describe on VeriGray,

and Fewshot on FAREBIO. After revision, rankings shift substantially. With **GPT-5**, DESCRIBE moves from rank 4 to rank 1 on STORYSUMM (71.4 → 77.5) and FAREBIO (76.0 → 87.3), while maintaining rank 1 on VERIGRAY. Most methods improve after revision, suggesting that corrected annotations yield more accurate and consistent evaluation, in line with Seo et al. (2025).

DESCRIBE benefits the most as the revisions primarily address discourse-level errors (e.g., fabricated relations, hedging removal, and misattached modifiers), which DESCRIBE is designed to detect. Under the original annotations, correct predictions on such cases were often penalized, suppressing its measured performance. Correcting these labels removes this penalty and reveals previously obscured differences, a pattern that holds across backbones but is most pronounced with `GPT-5`.

A potential concern is circularity, as DESCRIBE with `GPT-5` was used to flag candidate cases. However, flagging only determined which cases were reviewed. Revised labels were established through human inspection of the source evidence and verified by three independent LLMs (Stages 2–3), rather than by simply adopting DESCRIBE's predictions. Moreover, the revisions also do not uniformly benefit DESCRIBE (e.g., it remains fourth with `Gemini-3-Flash` on FAREBIO). These observations suggest that the revisions primarily correct discourse-level annotation errors rather than favoring a specific method.

8 Recommendations

Our findings align with prior work questioning the reliability of gold-standard annotations (Laban et al., 2023; Seo et al., 2025) and further identify a systematic pattern: in long-form summaries, annotation errors concentrate at the level of discourse relations rather than standalone individual facts. We distill four recommendations, ordered from post-hoc corrections to annotation design.

Recommendation 1: Use discourse-aware tools for post-annotation quality assurance. Our pipeline uncovers inconsistencies missed by both human annotators and existing evaluation systems. The key enabler is not model accuracy alone, but discourse-grounded rationales that make disagreements interpretable for human review. We therefore recommend applying discourse-aware evaluation tools as a post-annotation check, prioritizing manual review of cases with label disagreements. This would be particularly valuable for *fabricated relational links* and *scope and attachment* errors (§6), where individual facts are present in the source but their relations are not supported.

Recommendation 2: Reconsider unanimity-based aggregation. In STORYSUMM, 21 of 28 confirmed revisions were false negatives—sentences incorrectly labeled as faithful—often flagged by a single annotator whose annotation was overridden by unanimity (§5). For discourse-rich summaries, where relational errors are inherently harder to detect, unanimity can suppress valid minority judgments. We recommend explicitly reviewing dissenting annotations to better balance precision and recall. More broadly, given the subjectivity inherent in genre-dependent datasets (Subbiah et al., 2025), datasets should retain annotator rationales and explore methods that model label variation (Plank, 2022; Weber-Genzel et al., 2024).

Recommendation 3: Decompose claims before annotation. Our error analysis shows that annotators reliably verify individual facts but often miss the relationships between them, as reflected in frequent *semantic reframing* and *hedging removal* errors (§6). Our human verification shows that presenting pre-decomposed atomic claims alongside their discourse relations can reduce annotators' cognitive load and improve detection. This extends prior work on fine-grained annotation (Krishna et al., 2023) to a more linguistically grounded level by explicitly separating propositional content from epistemic and evaluative framing.

Recommendation 4: Incorporate relation-level categories. Many inconsistencies arise at the relation level—within sentences (misattached modifiers, dropped hedges) or across sentences (e.g., unsupported causal or temporal links)—yet all three benchmarks operate primarily at the sentence level, often with binary labels. While VERIGRAY's finer taxonomy yields fewer revisions (3.4%), it still lacks explicit relation-level categories such as *fabricated causality* or *temporal reordering*. We recommend augmenting sentence-leval labels with annotations of discourse relations (e.g., causal, temporal, attributive, evaluative) to capture whether these links are supported by the source.

9 Conclusion

We present a reassessment of three faithfulness benchmarks for long-form summaries, showing

that 3.4–5.4% of sentence-level annotations exhibit discourse-level inconsistencies. We introduce a five-category taxonomy—*fabricated relational links*, *scope and attachment errors*, *temporal reordering*, *hedging removal*, and *semantic reframing*—that explains how such errors arise and how a discourse-aware tool can detect them. Error patterns vary by genre and annotation design: narrative texts show more temporal and relational errors, while scientific and news texts are dominated by semantic reframing and hedging removal. Finer-grained taxonomies reduce but do not eliminate inconsistencies, and domain expertise can introduce biases through reliance on background knowledge. We propose four recommendations for future annotation efforts and release our revised labels for more reliable benchmark development.

Limitations

Our findings are limited to English-language summaries. Generalization to other domains and languages can be a promising future direction. In addition, error analysis and categorization were primarily conducted by a single annotator. To mitigate potential subjectivity, revisions are grounded in explicit textual evidence and supported by structured, discourse-based rationales, enabling transparent verification. We further validated our STORYSUMM corrections against the original annotators' free-text explanations, finding that 15 of 28 cases were independently flagged by at least one crowd worker. Finally, the auditing tool depends on discourse parsing quality, and errors in RST extraction may propagate to faithfulness judgments. More broadly, faithfulness evaluation remains inherently subjective at fine granularity, and revised human rationales may reflect biases of the underlying LLMs. To assess adjudication reliability directly, we had a second evaluator independently re-adjudicate a random sample of 15 flagged cases from the three datasets, reaching full agreement.

Ethical Considerations

The datasets used in this paper are all publicly available for research purposes. We acknowledge the potential for bias in human evaluation.

Responsible Disclosure. We contacted the authors of all three datasets prior to publication. The authors of STORYSUMM (Subbiah et al., 2024) confirmed several issues and shared raw annotation files to support our analysis. We release our revised labels alongside this paper to support transparent and reproducible benchmark development, rather than to undermine prior annotation efforts. The authors of VERIGRAY (Ding et al., 2025) updated the annotations based on our feedback, demonstrating the practical impact of our findings. The authors of FAREBIO (Fang et al., 2024) kindly provided the unprocessed annotation files but were unable to revise the dataset due to limited access to domain experts. Further validation with domain experts would strengthen revisions to FAREBIO and remains future work.

Annotator Attribution. We attribute observed errors to annotation protocol design (e.g., aggregation rules, taxonomy granularity) rather than individual annotators. We emphasize that the original annotators—crowd workers, graduate students, and medical doctors—operated within given guidelines; the identified issues reflect *procedural* limitations, not annotator competence.

Intended Use. Our revised labels and error taxonomy are intended to study the impacts of discourse-level errors in long-form summary evaluation and inform future annotation designs and practices. They are not meant to discredit existing benchmarks, as the vast majority of annotations (94.6–96.6%) remain unchanged after our reassessment.

Acknowledgments

This research was supported in part by the University of Pittsburgh Center for Research Computing and Data, RRID:SCR_022735, through the H2P cluster, which is supported by NSF award number OAC-2117681. We want to thank the members of the Pitt PETAL group and anonymous reviewers for their valuable suggestions, which helped improve this work. We also thank Zhehan Tiffany Zhu for her contributions to and discussions of the human evaluation.

References

Qiang Ding, Lvzhou Luo, Yixuan Cao, and Ping Luo. 2025. The gray zone of faithfulness: Taming ambiguity in unfaithfulness detection. *arXiv preprint arXiv:2510.21118*.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Biaoyan Fang, Xiang Dai, and Sarvnaz Karimi. 2024. Understanding faithfulness and reasoning of large language models on plain biomedical summaries. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9890–9911, Miami, Florida, USA. Association for Computational Linguistics.

Gérard Genette. 1983. *Narrative discourse : an essay in method / Gérardrard Genette ; translated by Jane E. Lewin ; foreword by Jonathan Culler.* Cornell paperbacks. Cornell University Press, Ithaca, N.Y.

Ameya Godbole and Robin Jia. 2025. Verify with caution: The pitfalls of relying on imperfect factuality metrics. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22889–22912, Vienna, Austria. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. SNaC: Coherence error detection for narrative summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 444–463, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sil Hamilton, Matthew Wilkens, and Andrew Piper. 2025. Narrabench: A comprehensive framework for narrative benchmarking. *Preprint*, arXiv:2510.09869.

Huan Yee Koh, Jiaxin Ju, He Zhang, Ming Liu, and Shirui Pan. 2022. How far are we from robust long abstractive summarization? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2682–2698, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Philippe Laban, Wojciech Kryscinski, Divyansh Agarwal, Alexander Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. 2023. SummEdits: Measuring LLM ability at factual reasoning through the lens of summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9662–9676, Singapore. Association for Computational Linguistics.

Yuho Lee, Taewon Yun, Jason Cai, Hang Su, and Hwanjun Song. 2024. UniSumEval: Towards unified, fine-grained, multi-dimensional summarization evaluation for LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3941–3960, Miami, Florida, USA. Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wooseok Seo, Seungju Han, Jaehun Jung, Benjamin Newman, Seungwon Lim, Seungbeen Lee, Ximing Lu, Yejin Choi, and Youngjae Yu. 2025. Verifying the verifiers: Unveiling pitfalls and potentials in fact verifiers. In *Second Conference on Language Modeling*.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. FineSurE: Fine-grained summarization evaluation using LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 906–922, Bangkok, Thailand. Association for Computational Linguistics.

Melanie Subbiah, Faisal Ladhak, Akankshya Mishra, Griffin Thomas Adams, Lydia Chilton, and Kathleen McKeown. 2024. STORYSUMM: Evaluating faithfulness in story summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9988–10005, Miami, Florida, USA. Association for Computational Linguistics.

Melanie Subbiah, Akankshya Mishra, Grace Kim, Liyan Tang, Greg Durrett, and Kathleen McKeown. 2025. Is the top still spinning? evaluating subjectivity in narrative understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 185–203, Suzhou, China. 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.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

Kun Zhang, Oana Balalau, and Ioana Manolescu. 2025. Structured discourse representation for factual consistency verification. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 820–838, Vienna, Austria. Association for Computational Linguistics.

Yang Zhong and Diane Litman. 2025a. Discourse-driven evaluation: Unveiling factual inconsistency in long document summarization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2050–2073, Albuquerque, New Mexico. Association for Computational Linguistics.

Yang Zhong and Diane Litman. 2025b. A tale of evaluating factual consistency: Case study on long document summarization evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12511–12532, Vienna, Austria. Association for Computational Linguistics.

A Benchmark Details

We include additional details of the three studied benchmarks in the sections below.

A.1 VERIGRAY

The error taxonomy of VERIGRAY is included in Table 5. Following the original paper's suggestions, we exclude the ambiguous and "no-fact" cases in the experiments presented in §7, as these changes are hard to categorize and assess.

A.2 STORYSUMM

The authors ask three crowd-sourcing annotators to assess whether each summary sentence is consistent with the original story. To reduce the influence of subjective "commentary" sentences (e.g., "*The story reflects the enduring bonds of friendship ...*" which interpret story themes rather than describe story plot), the annotation interface provides an additional "N/A, just commentary" label in addition to the "yes" or "no" options. The initial inter-annotator agreement reaches an almost perfect agreement of Fleiss' kappa of 0.85.

Recognizing the challenges of evaluating narrative summaries, the authors introduce two additional procedures. In the *Expert* setting, three authors adjudicate summary-level labels. In the *Hybrid* setting, the authors first employ GPT-4 to generate potential inconsistencies between the source document and the summary, after which a new group of crowd workers reassesses the summaries using these suggestions. The final dataset merges unfaithful labels agreed upon by all three original annotators with adjudicated expert labels and labels produced through this hybrid human–AI process.

Despite these quality control steps, some aspects of the annotation scheme remain unclear, as the author only released a final version. In particular, crowd workers could assign an *"N/A, Commentary"* label to subjective sentences, but how these labels are incorporated into the final binary scheme is not documented. To examine this dataset, we obtained the raw per-annotator files from the original authors, including individual votes and free-text explanations, which allow us to analyze the potential impact of these ambiguous cases.

While conducting small-scale error analysis on the STORYSUMM predictions, we notice that several LLM-based approaches consistently reached different predictions compared to the sentence-level labels in the original dataset. These discrepancies are confirmed to be either incorrect labels or subjectively annotated cases (where readers can have different interpretations given the vagueness of presentation), as reported in the same authors' follow-up work (Subbiah et al., 2025).

We contacted the original authors of the STORYSUMM paper to validate our hypothesis that a certain level of annotation error exists in the original dataset. The authors confirmed that they collected three crowd-sourced annotator labels for each sentence and assigned the final label as "No" (unfaithful) only when all three annotators agreed; they also willingly shared the raw annotation files. We conduct the following analyses to validate the sentence-level labels, which have been overlooked in the original paper (they only apply diverse evaluation protocols on the summary-level).

Ignoring Commentary Sentences. Prior work analyzed STORYSUMM's annotation files, finding that commentary sentences (at least one out of three annotators) are included in the released dataset. To align with the original paper's definition, we exclude the same 19 sentences through from both summary-level and sentence-level assessments to maintain consistency with the evaluation guidelines of the original paper.

A.3 FAREBIO

This dataset contains 175 LLM-generated summaries ($1,445$ sentences) of biomedical research abstracts and introductions, annotated by two medical doctors along two binary dimensions: faithfulness and factual hallucination, with highlighted supporting evidence spans. The reported inter-annotator agreement (Cohen's Kappa = 0.48 on 34 double-annotated sentences) itself signals the difficulty of faithfulness judgment in this domain. While the binary scheme simplifies the annotation task, the limited disclosure regarding annotator training beyond the 34 example sentences and labeling guidelines, together with the reliance on single annotations for the majority of the dataset, leaves room for quality analysis.

We contacted the original authors and obtained a zip file containing the original annotation data. However, we were unable to recover the per-annotator annotation details due to the absence of documentation and the presence of complex annotation cache files.

Document	Target Sentence	Annotation
. . . qatar captial doha , home to the aspire dome , beat eugene to host the 2019 event in granting the championships to eugene the iaaf council . . .	Eugene had previously failed in its bid to host the 2019 event, which was awarded to Doha.	`Explicitly-Supported`.
nicklaus holds up his ball to an adoring crowd as gary player (left) and ben crenshaw salute the great crenshaw and nicklaus fist pump following his ace on the 130-yard	This feat was witnessed by fellow golfers Gary Player and Ben Crenshaw.	`Implicitly-Supported`. **Reason:** The document supports the claim but does not explicitly mention the highlighted text.
A Gareth Anscombe drop-goal edged Blues 23-20 ahead after Gloucester Josh Hohneck was yellow carded. But unanswered second-half tries from Jonny May, Marshall, Mark Atkinson and Henry Purdy sealed Gloucester's win.	The Gloucester Rugby team won 23-20 against Cardiff Blues in a European Cup competition match.	`Fabricated`. **Reason:** The 23-20 is not the final score.
Italy's National Institute of Geophysics and Volcanology (INGV) said the quake struck at 15:48 (14:48 GMT), with its epicentre in Garfagnana.	The quake struck at 15:48 GMT and was followed by several aftershocks	`Contradicting`. **Reason:** It should be 14:48 GMT.
ormer Russian FSB colonel Igor Girkin, also known as Strelkov, who was then a key rebel commander in eastern Ukraine. Access to Anonymous International's website is currently blocked in Russia.	3. **Anonymous International (Shaltay Boltay)** – Known for leaking Kremlin documents, the group has also released material on Ukraine, including emails from a Russian rebel commander.	`Ambiguous`. **Reason:** "russian rebel commander" is ambiguous: (russian rebel) commander vs. russian (rebel commander)
Francis I (Franz Stefan , Francois Etienne 8 December 1708 – 18 August 1765) was Holy Roman Emperor and Grand Duke of Tuscany , though his wife effectively executed the real powers of those positions	The passage distinguishes between two historical figures named Francis I—one a Habsburg emperor and the other a Valois king of France—while focusing primarily on the latter's reign, rivalries, and military strategies.",	`Out-Dependent` **Reason**: Habsburg is not mentioned. However, `https://en.wikipedia.org/wiki/Francis_I,_Holy_Roman_Emperor` writes that "Following the death of his father-in-law, Charles VI, in 1740, Francis and Maria Theresa became the rulers of the Habsburg domains."
a pit crew member was hit by a car on sunday during the inaugural indycar grand prix of louisiana .	The incident occurred at the NOLA Motorsports Park in Avondale. (Word count: 99) me know if you'd like any adjustments!	`No-Fact`. **Reason**: This is considered as a meta note generated by LLMs, which is irrelevant to the summary and should be excluded from evaluation.

Table 5: **Examples of the taxonomy defined in VERIGRAY** (Ding et al., 2025). The key segment that drives the annotation decision for each example is in red. We select these examples from the dataset.

B Prompts

We include the prompts for FineSURE (Song et al., 2024) and a modified version of VG-prompt (Ding et al., 2025) in Table 6, which are adopted from the corresponding papers and codebases. For Fewshot, ClearCheck, and FaithLens, we used the model checkpoints open-sourced by the original authors and used their released prompts to process our data. For DESCRIBE, we use the publicly available tool.

C Label Revision Experiments

Backbones. We picked two closed-source models, `GPT-5`[7] and `Gemini-3-Flash`[8], and used their official APIs for our experiments. For GPT-5, we set the reasoning effort to medium due to budget constraints, while for `Gemini-3-Flash`, we used high reasoning effort. Due to computational cost, all results are averaged over two runs .

Baselines. We include the prompt for FineSURE (Song et al., 2024), a modified version of VG-prompt (Ding et al., 2025) in Table 6, adopted from the corresponding papers and codebases. For Fewshot, we follow the `FEW_SHOT_TEMPLATE` in the official codebase.[9] For DESCRIBE, we adopt the publicly available codebase.

Datasets. For the original datasets (prior to label revision), we experiment with the following versions:

- STORYSUMM: `https://github.com/melaniesubbiah/storysumm/blob/main/storysumm_w_subj.json`, retrieved in mid-February.
- VERIGRAY: `https://huggingface.co/datasets/Ding-Qiang/veri-gray-20251007`.
- FAREBIO: retrieved from `https://data.csiro.au/collection/csiro:63362`.

For the revised datasets:

- STORYSUMM: We correct the 28 reassessed labels and further exclude 19 "commentary

[7] API name: `gpt-5-2025-08-07`

[8] API name: `gemini-3-flash-preview`

[9] `https://github.com/just1nseo/verifying-the-verifiers/blob/main/data/templates.py`

FineSURE:

You will receive a transcript followed by a corresponding summary. Your task is to assess the factuality of each summary sentence across the following nine categories:
no error: the statement aligns explicitly with the content of the transcript and is factually consistent with it.

out-of-context error: the statement contains information not present in the transcript.
entity error: the primary arguments (or their attributes) of the predicate are wrong.
predicate error: the predicate in the summary statement is inconsistent with the transcript.
circumstantial error: the additional information (e.g., location or time) specifying the circumstance around a predicate is wrong.
grammatical error: the grammar of the sentence is so incorrect that it becomes meaningless.
coreference error: a pronoun or reference has a wrong or non-existing antecedent.
linking error: incorrect discourse linkage between statements (e.g., temporal or causal relations).
other error: the statement contains any factuality error not defined above.
Instruction:

First, compare each summary sentence with the transcript.
Second, provide a single sentence explaining the factuality assessment.
Third, assign exactly one error category for each sentence.
Provide your answer in JSON format. The answer should be a list of dictionaries whose keys are "sentence", "reason", and "category":
[{"sentence": "first sentence", "reason": "your reason", "category": "no error"},
{"sentence": "second sentence", "reason": "your reason", "category": "entity error"},
{"sentence": "third sentence", "reason": "your reason", "category": "out-of-context error"}, ...]

Transcript:
{DOCUMENT}

Summary with N sentences:
1. {summary sentence 1}
2. {summary sentence 2}
...
N. {summary sentence N}

VG-prompt

You are judging the faithfulness of each sentence of a summary to the source document. The faithfulness labels should be selected from the following options:
A. Explicitly-Supported: all atomic facts of the sentence appear verbatim (up to lexical or syntactic transformation) within the document.
B. Generally-Supported: the document entails the sentence, but the sentence is not explicitly supported. Minor differences are allowed only if the sentence adopts a weaker or less certain claim than the document. If any part of the sentence adopts a stronger or more certain claim, select Fabricated.
C. Inconsistent: the sentence logically contradicts the document.
D. Fabricated: the sentence does not contradict the document, but is neither implied by the document nor external world knowledge.
E. Out-Dependent: the sentence is not implied by the document alone, but is implied by combining the document with external world knowledge.
F. Ambiguous: the sentence or the document admits multiple interpretations.
G. No-Fact: the sentence does not contain factual content.
Instruction:
Assess each summary sentence independently.
The output should be a JSON list enclosed within the special tag `<FINAL_PRED></FINAL_PRED>`. The list must contain one dictionary per sentence, with keys "sent_id", "sentence", "rationale", and "faithfulness_label".
Example output format:
`<FINAL_PRED>`
`[{"sent_id": 1, "sentence": "...", "rationale": "...", "faithfulness_label": "A"}, ...]`
`</FINAL_PRED>`
Ensure the number of output entries matches the number of summary sentences exactly.

Document:
{DOCUMENT}

Summary with N sentences:
1. {summary sentence 1}
2. {summary sentence 2}
...
N. {summary sentence N}

Table 6: FineSURE and VG-prompt prompts for faithfulness evaluation.

sentences" based on the analysis presented in Appendix A.2.

- VERIGRAY: Our findings were incorporated by the original authors, resulting in an updated release with additional author-side inspection. We therefore use the official updated version, `20251225.jsonl`.
- FAREBIO: Because the authors were unable to review or validate the proposed revisions, we apply all our proposed corrections to the original dataset and use this modified version to reassess model performance.

Evaluation Setup. For VERIGRAY (Ding et al., 2025), we follow their first protocol to compute the balanced accuracy. Following their setup, we remove the `Not Sure` classes (`Out-Dependent` and `Ambiguous`) and `No-Fact` cases. The remaining classes can then be merged into two categories. `Explicitly-Supported` and `Implicitly-Supported` are merged into the **Faithful Class**, while `Fabricated` and `Contradicting` constitute the **Unfaithful Class**. We further report the balanced accuracy. Predictions are aligned accordingly: for methods with fine-grained predictions that follow the VERIGRAY taxonomy, specifically VG-prompt and DESCRIBE, classes with a faithfulness degree not less than a threshold (defined below) are considered **Faithful**, with the rest as **Unfaithful**. Here, the *order of faithfulness degree* is defined as:

$$\begin{aligned} \text{Contradicting} < \text{Fabricated} < \text{Ambiguous} \\ < \text{No-Fact} < \text{Out-Dependent} \\ < \text{Implicitly-Supported} < \text{Explicitly-Supported} \end{aligned} \tag{1}$$

The threshold is set to the `Implicitly-Supported` class, following the original paper.