# DESCRIBE: Discourse-Driven Extraction and Structuring of Claim Verification for Better Faithfulness Evaluation

**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

Evaluating the faithfulness of multi-sentence summaries generated by language models remains challenging due to their length and the subtlety of factual errors. While claim decomposition methods have proven useful for assessing short summaries (1-3 sentences), it remains unclear whether existing approaches can reliably and consistently decompose and validate claims in longer summaries across diverse domains. To better understand this issue, we first conduct an empirical study analyzing the quality of claims produced by existing decomposition approaches. Motivated by our finding, we propose a new discourse-informed method, DESCRIBE, which first extracts claims from the full summary using discourse structure, and then evaluates them through a structure-guided process that progresses from intra- to inter-sentential claim validation, enabling more fine-grained faithfulness evaluation. Experimental results across benchmarks reveal the benefits of discourse-guided decomposition and evaluation. Further analysis of model-generated error rationales and ablation studies shows our method effectively identifies nuanced errors arising from complex discourse structures.[1]

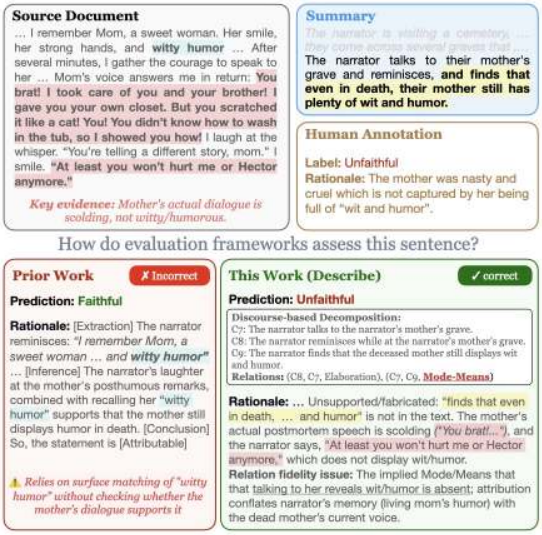



Figure 1: **DESCRIBE detects discourse-level faithfulness errors missed by prior work**. **Top left:** The source document contrasts the narrator's recollection of his mother's "witty humor" with her actual scolding dialogue. **Top right:** an LLM-generated summary claims she "still has plenty of wit and humor"; a human annotator labels the sentence *unfaithful* with rationales. **Bottom left:** prior work extracts surface-level claims and incorrectly label the sentence faithful. **Bottom right:** our work decomposes the summary into discourse-linked claims, checks the relations against the source, and correctly flags the unsupported characterization.

## 1 Introduction

Checking the faithfulness of summaries, i.e., ensuring the information in summaries is factually consistent with the source document (Maynez et al., 2020; Kryscinski et al., 2020; Cao and Wang, 2021; Krishna et al., 2023; Zhang et al., 2024) remains a challenging task. With the emergence of stronger large language models (LLMs, Yang et al. 2025; OpenAI et al. 2024), faithfulness errors have shifted from surface-level grammatical issues to more subtle, contextually nuanced ones (Lee et al., 2024; Zhong and Litman, 2025b). As fine-grained data inspection becomes more central to faithfulness evaluation, researchers have increasingly highlighted the inherent difficulties in labeling ambiguous cases, where the available evidence supports multiple plausible interpretations or involves subjectivity (Seo et al., 2025; Subbiah et al., 2025; Ding et al., 2025). These challenges highlight the limitations of existing annotation practices in producing reliable, fine-grained faithfulness labels beyond binary classification, motivating the development of new approaches for more interpretable evaluation across different faithfulness error types.

Previous work has adapted the *Decompose-Then-*

[*]Equal senior contribution.
[1]Codebase and parses are released in https://github.com/cs329yangzhong/faithfulEval4LongSumm.



*Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026*, pages 2737–2771
November 6-10, 2026 ©2026 Association for Computational Linguistics

*Verify* pipeline from fact-checking (Min et al., 2023) to evaluate short summaries, which breaks summaries into atomic claims and verifies them independently. While effective for data in the news domain (Liu et al., 2023d), this paradigm has fundamental limitations for assessing multi-sentence summaries or for other text types. Most approaches rely on simple few-shot prompts crafted in the biography domain (Scirè et al., 2024) for producing atomic claims (individual, verifiable unit of information) among individual summary sentences, leading to incomplete or missing facts (Liu et al., 2025). Crucially, this atomicity overlooks the discourse relations that bind sentences together. Furthermore, most existing fact-checkers overly rely on surface-level semantic matching and fail to generalize, even when trained on synthetic datasets containing reasoning-style data (Seo et al., 2025). As illustrated in Figure 1, such approaches can miss critical discourse-level faithfulness errors.

To address this gap, we propose DESCRIBE, a discourse-informed framework for evaluating the faithfulness of long-form summaries (four or more sentences, >100 words). We first show that existing decomposition-based methods struggle with multi-sentence summaries, often introducing context omission and meaning alteration errors (§3). Motivated by these findings, DESCRIBE uses the Rhetorical Structure Theory (RST, Mann and Thompson 1988) to preserve discourse structure during evaluation. Unlike prior work that relies on explicit discourse connectives (Zhang et al., 2025) and overlooks cross-sentence dependencies and implicit relations, DESCRIBE employs RST-based claim extraction to preserve intra- and inter-sentential relations, followed by a structured inference procedure that validates claims and their relations hierarchically from the local to document-level (§4). This shifts faithfulness evaluation from isolated, decomposed claims to jointly evaluating claims and the discourse relations that connect them.

Experiments on two benchmarks show that DESCRIBE consistently outperforms fine-tuned fact-checkers and prompt-based LLM baselines in sentence-level faithfulness assessment (§6.1). Further analyses demonstrate that both proposed components contribute meaningfully to downstream task performance (§6.2). In addition, DESCRIBE produces rationales that more closely align with human reasoning and exhibit stronger discourse awareness according to both automatic metrics and human evaluation (§6.3).

## 2 Related Work

**Claim Decomposition in Summarization.** Prior work has shown that claim decomposition and decontextualization (Choi et al., 2021) may not benefit the claim evaluation task (Tang et al., 2024). Hu et al. (2025) observe benefits only for complex inputs, and finer-grained decomposition can harm the performance. In long-form factuality checking, VeriFact (Liu et al., 2025) highlights issues of incomplete and missing facts from few-shot prompting templates designed in the biography domain (Min et al., 2023). While claim decomposition has proven effective for short summaries (1-3 sentences) (Wan et al., 2024; Scirè et al., 2024; Yang et al., 2024), assessing faithfulness in longer summaries (over four sentences) poses distinct challenges (Zhong and Litman, 2025b), including complex discourse structure, cross-sentence dependencies, and subtle meaning distortions. *We introduce a discourse-guided framework that explicitly models document-level coherence to enable robust validation of both intra- and inter-sentential content via claim decomposition and relation modeling.*

**Discourse-inspired Approaches for Faithfulness Evaluation.** Recent work has incorporated discourse insights to improve faithfulness evaluation in abstractive summarization beyond isolated claim verification. Prior work in extractive summarization (Zhang et al., 2023) shows that unfaithfulness often arises from missing or incorrect discourse connections, motivating subsequent efforts to adapt these ideas to abstractive summarization. Existing approaches primarily use shallow discourse cues or local relational structures, such as modeling discourse markers based on abstraction meaning representation (AMR, Lee et al. 2025), or enriching atomic fact tuples (Zhang et al., 2025) with discourse connectives from the Penn Discourse Tree Bank (PDTB, Prasad et al. 2019). Zhong and Litman (2025a) adopt sentence-level discourse patterns to adjust each sentence to a summary-level assessment. We further discuss distinctions between DESCRIBE and prior work in Appendix A.

More recently, Verifact (Liu et al., 2025) introduces a multi-stage framework that exploits inter-sentential PDTB relations to recover missing information in extracted claims. However, it only relies on shallow discourse representations and often compromises claim atomicity by merging claims. *In contrast, our approach focuses specifically on faithfulness evaluation for abstractive summaries*

| | Text Diversity | Human Annotation | | | | | Dataset Statistics | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Doc.Src | Granularity | Taxonomy | Annot. Dimension | Method | w/ rationale | # Summary | # Sent | Doc.Word | Sum.Sent | Sum.Word |
| STORYSUMM | narratives | sent-level | binary | faithful+subjective | auto+expert | Yes | 96 | 580 | 767 | 6 | 139 |
| VERIGRAY | news | sent-level | fine | faithful | auto+expert | Yes | 412 | 2,044 | 428 | 5 | 122 |

Table 1: Dataset statistics for STORYSUMM (Subbiah et al., 2024) and VERIGRAY (Ding et al., 2025): input source (Doc.Src), annotation granularity, number of annotated test cases (# Summary, # Sentences), average document length (Doc.Word), and average summary length in sentences (Sum.Sent) and words (Sum.Word).

*by using RST to model document-level intra- and inter-sentential relations. This enables a structured inference procedure that effectively identifies faithfulness errors overlooked by prior work.*

## 3 Analyzing Decomposition Approaches

As more powerful LLMs emerge, faithfulness errors in model-generated summaries have shifted from primarily surface-level grammatical or disfluency issues to more subtle and context-dependent errors (Lee et al., 2024; Zhong and Litman, 2025b). To study this, we use two benchmarks with human-annotated LLM-generated summaries: STORYSUMM (Subbiah et al., 2024) and VERIGRAY (Ding et al., 2025). Table 1 presents details.

**Decomposition Baselines and Evaluations.** We analyze two representative claim decomposition approaches that incorporate discourse information to maintain cross-sentence coherence. Specifically, we consider **FABLES** (Kim et al., 2024), which extracts atomic claims while preserving temporal, locational, and causal context,[2] and **VERIFACT**, a multi-stage framework that uses discourse-based prompts to detect and refine incomplete and missing facts (Liu et al., 2025). We examine the average number of sub-claims after decomposition and the relative ratio of sub-claims to sentences, and further analyze decomposition errors using a prompt-based automatic framework with the few-shot evaluation prompts from Hu et al. (2025),[3] which categorized errors as (1) omission of context information (missing logical relations or core claims), (2) ambiguity, (3) alteration of original meaning, and (4) over-decomposition (excessive fragmentation of the claim into redundant sub-claims).

**Findings.** We compare the number of claims generated by each method. On average, FABLES produces 12.0 claims per text, while VERIFACT generates 19.3. Its higher claim-per-sentence ratio (5.1 vs. 2.6) indicates over-decomposition. Figure 2 shows that over 40% of decomposed outputs contain errors, notably *over-decomposition*, *omission of context information*, and *alteration of original meaning*, especially in summaries labeled unfaithful. These results highlight that existing discourse-inspired decomposition methods often disrupt semantic content and discourse structure, underscoring the need for a better discourse-aware extraction method that preserves contextual dependencies.

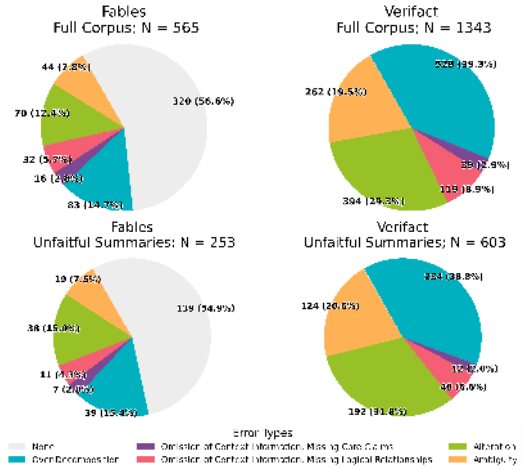



Figure 2: Decomposition error distribution across approaches: full corpus (top) vs. summaries with $\geq 1$ human-labeled error (bottom).

## 4 OUR APPROACH: DESCRIBE

We present DESCRIBE (**D**iscourse-driven **E**xtraction and **S**tructuring of **C**laim **R**eason**I**ng for **B**etter-grounded Faithfulness **E**valuation), a framework that incorporates discourse structure for fine-grained faithfulness evaluation. Unlike prior approaches that only assess individual claims in isolation, DESCRIBE explicitly models intra- and inter-sentential relations alongside dimensions such as claim correctness, enabling detection of more subtle causation and attribution errors.

To preserve these relations during decomposition and evaluation, we use Rhetorical Structure

[2]The original paper reports 100% precision for decomposed claims in book summaries.

[3]Manual review of 90 sampled instances from both datasets yielded a 92% agreement rate on detected errors.

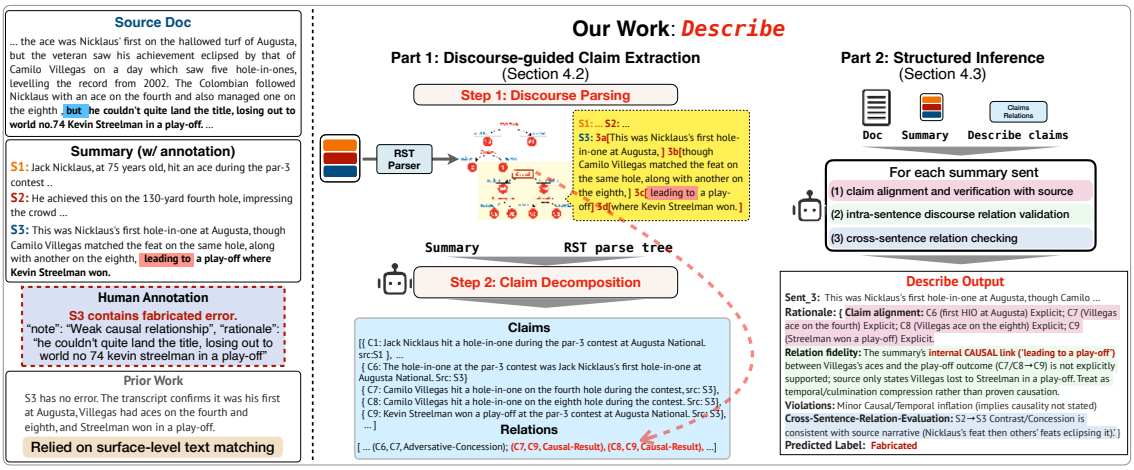



Figure 3: **Overview of DESCRIBE (illustrated with a VERIGRAY example). Left**: Source document and summary with human annotations; prior work misses the error in S3 by checking individual facts without considering relational structure. **Part 1** (§4.2): Discourse-guided claim extraction parses the summary into an RST tree and identifies intra- and inter-sentential discourse relations. **Part 2** (§4.3): Structured inference validates each summary sentence by aligning claims to the source and checking intra- and cross-sentence discourse relations.

Theory (RST, Mann and Thompson 1988), which represents text as a hierarchical tree of elementary discourse units (EDUs), typically short clauses, connected by relations such as ELABORATION, CAUSAL, CONTRAST, and ATTRIBUTION. This structure captures both local (intra-sentential) and global (inter-sentential) dependencies that may otherwise be lost during claim decomposition. We adopt the RST relation inventory from the multi-genre corpus, GUM (Zeldes, 2017), and train a parser, DMRST (Liu et al., 2021) on GUM v11. Implementation details and parser performance are included in Appendix C. Figure 3 illustrates the overall approach, and the middle panel shows a parsed RST-style discourse tree.

### 4.1 Problem Formulation

Given a source document $D$ and a candidate summary $S = \{s_1, s_2, ..., s_n\}$ consisting of $n$ sentences, the goal is to produce sentence-level faithfulness labels $y = \{y_1, y_2, ..., y_n\}$ along with interpretable rationales $r = \{r_1, r_2, ...r_n\}$ justifying each prediction. We adopt a fine-grained faithfulness taxonomy (Ding et al., 2025), consisting of seven categories: *Explicitly-Supported*, *Generally-Supported*, *Contradicting*, *Fabricated*, *Out-Dependent*, *Ambiguous*, and *No-Fact*.

### 4.2 Discourse-Guided Claim Extraction

We use an RST-guided prompting strategy to guide an LLM for discourse-aware claim decomposition and relation extraction. Given a Summary $S$ and an optional RST parse produced by a trained parser, the model breaks each sentence into fully self-contained atomic claims (Part 1 in Figure 3), explicitly resolving anaphora and retaining temporal, locational, and causal modifiers for contextual completeness, following Kim et al. (2024). To capture document-level coherence, the model further identifies intra- and inter-sentential relations using the parser-predicted RST tree when available and infers relations among claims within and across sentences. While predicted RST trees may contain errors (e.g., in Step 1, EDU 3cd is incorrectly attached to 3ab in Figure 3), core discourse relations like CAUSAL remain largely intact.[4] The final output is a JSON-formatted structured claim set with a directed discourse relation graph encoding source and target claims, RST relation types, and intra-/inter-sentential scope, yielding a discourse-preserving summary representation for downstream structure-faithfulness reasoning.

### 4.3 Structured Inference

Given the source document, a candidate summary, and the discourse-guided claim decomposition from §4.2, we prompt an LLM to perform multi-dimensional verification of sentences within the summary. The model first aligns each summary sentence to its corresponding atomic claims and

[4]Additional analysis in Appendix G.4 shows that zero-shot prompted LLM to parse the sentence leads to performance degradation on the downstream task, potentially attributed to the weaker performance as compared to the trained parser (Liu et al., 2026).

searches the source text for explicit or entailed evidence that supports each claim. It then validates the discourse relations given the source document, assessing the fidelity of temporal ordering, causal structure, attribution, and rhetorical relations. For instance, Figure 3 shows our approach validates the causal relation between Claims 7 and 9 by assessing if *"the Aces made by Camilo"* serves as the cause for the model's categorization of faithfulness violations across predefined evaluative dimensions.

Beyond the within-sentence claim and relation validations, the model additionally performs cross-sentence relation evaluation to detect potential long-range discourse transition inconsistencies. Lastly, we aggregate claim-level and relation-level evidence to assign one of the seven fine-grained labels described above that are defined in Ding et al. (2025) to each summary sentence, producing a structured JSON output with explicit rationales to support the final decision.

Combining both modules, our framework DESCRIBE operates in a zero-shot setting, requiring no task-specific training data. This design reflects both a practical constraint, as no existing benchmark provides relation-level faithfulness supervision, and a deliberate goal of portability across backbone LLMs of different scales and architectures. We include implementation details in Appendix D.

## 5 Experimental Setup

We include three groups of faithfulness evaluation baselines alongside our proposed approaches: **(1) Fine-tuned fact-checkers**, selected as recent and widely adopted methods for grounded faithfulness evaluation, which are powered mainly by natural language inference models trained over synthetic data on fact-checking tasks. Specifically, we include **MiniCheck** (Tang et al., 2024), **ClearCheck** (Seo et al., 2025), and **FaithLens** (Si et al., 2026); **(2) Discourse-inspired approaches**: we experiment with **StructScore** (Zhong and Litman, 2025a), which leverages the summaries' discourse structure information to adjust sentence-level assessments, and **FDSpotter** (Zhang et al., 2025), which decomposes the long summaries into atomic fact tuples empowered with discourse relations and further verifies each tuple; **(3) Prompt-based approaches** were carried out with two open-sourced LLMs, `Qwen3-30B-A3B-Thinking-2507` (Yang et al., 2025) and `gpt-oss-120B` (OpenAI, 2025), as well as `GPT-5` and `Gemini-3-Flash`.

Regarding prompt choices, we include three varieties. **FineSURE** (Song et al., 2024) categorizes sentences by fine-grained error types (Pagnoni et al., 2021); **Fewshot** prompt (Seo et al., 2025) provides few-shot claim supportiveness assessment examples; and **VG-prompt** is a modified version of the original VERIGRAY prompt (Ding et al., 2025). For **DESCRIBE**, we use the same backbone for both modules, and further include **DESCRIBE**$^{\text{gpt5}}$, which incorporates superior discourse-guided claims generated by `GPT-5` as input for structured inference. We report the average of three runs for open-source models and two runs for `GPT-5` and `Gemini-3-Flash` due to computational cost. Details are provided in Appendix E.

**Prediction Evaluation Metrics.** For both benchmarks, we report balanced accuracy (BAcc, Brodersen et al. 2010) and Matthews correlation coefficient (MCC, Matthews 1975) as the binary classification metrics. For STORYSUMM, we also report Cohen's Kappa (Artstein and Poesio, 2008) on the summary-level and report all three on the sentence-level. For VERIGRAY, following Ding et al. (2025), we further report hallucination detection recall, precision, and F1 (i.e., recall, precision, and F1 for the Unfaithful class), merging the fine-grained labels into the binary setting as described in the original work. We also report the ranking loss to make use of all fine-grained annotations. A full description of these metrics is provided in Appendix E.4.

## 6 Results

Below we report DESCRIBE's performance against specialized fine-tuned fact-checkers and prompt-based approaches on two benchmarks.

### 6.1 Long-Form Summary Faithfulness

Table 2 reveals that current faithfulness evaluation methods exhibit notable limitations when applied to long-form summaries generated by recent LLMs.

**Fine-tuned fact checkers underperform.** Specialized checkers (ROWS 1-3) consistently underperform prompt-based approaches at the sentence level across the board, despite being trained on large-scale synthetic datasets. Methods such as FaithLens achieve only moderate balanced accuracy (below 60% on STORYSUMM), while the consistently low hallucination recall on VERIGRAY indicates difficulties in reliably identifying missing or distorted information within extended narrative structure or nuanced errors.

| ID | Method | StorySumm summary-level Coh.-k↑ | StorySumm summary-level BAcc.↑ | StorySumm summary-level MCC↑ | StorySumm sent-level Coh.-k↑ | StorySumm sent-level BAcc.↑ | StorySumm sent-level MCC↑ | VeriGray sent-level BAcc.↑ | VeriGray sent-level MCC↑ | VeriGray sent-level Hal. R.↑ | VeriGray sent-level Hal. P.↑ | VeriGray sent-level Hal. F1↑ | VeriGray sent-level $L_{rank}$ ↓ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ***Fine-tuned fact checker*** | | | | | | | | | | | | |
| 1 | MiniCheck | 0.07 | 53.6 | 0.07 | 0.18 | **58.1** | 0.18 | **59.6** | 0.23 | **26.3** | 42.2 | **32.4** | **42.8** |
| 2 | ClearCheck | **0.21** | **62.9** | **0.30** | 0.15 | 55.3 | 0.19 | 57.7 | **0.28** | 17.3 | **65.2** | 27.3 | 44.7 |
| 3 | FaithLens | 0.08 | 54.7 | 0.09 | **0.19** | 57.7 | **0.20** | 59.1 | 0.27 | 22.0 | 52.8 | 31.1 | 43.3 |
| | ***Discourse-inspired Approaches*** | | | | | | | | | | | | |
| 4 | StructScore | **0.25** | 61.1 | **0.29** | **0.24** | **67.5** | **0.27** | **64.4** | **0.26** | **45.0** | 35.3 | **39.6** | **39.8** |
| 5 | FDSpotter | **0.25** | **64.2** | 0.28 | 0.20 | 58.6 | 0.20 | 54.5 | 0.15 | 12.9 | **38.1** | 19.3 | 46.9 |
| | ***LLM-based approaches with different backbones*** | | | | | | | | | | | | |
| | `Qwen3-30B-A3B` | | | | | | | | | | | | |
| 6 | Fewshot | **0.24** | **63.2** | **0.26** | 0.27 | 62.0 | 0.28 | 70.7 | 0.49 | 45.7 | 67.8 | 54.6 | 38.1 |
| 7 | FineSURE | 0.14 | 58.2 | 0.16 | **0.32** | 63.4 | **0.33** | 69.8 | 0.49 | 43.1 | 70.6 | 53.5 | 38.7 |
| 8 | VG-prompt | 0.17 | 59.3 | 0.18 | 0.27 | 62.0 | 0.28 | 73.2 | **0.56** | 49.3 | **77.3** | **60.1** | 36.6 |
| 9 | **DESCRIBE** | 0.22 | 60.9 | 0.22 | 0.26 | 62.8 | 0.26 | 74.4 | 0.52 | **54.7** | 65.1 | 59.4 | **36.1** |
| 10 | **DESCRIBE**$^{gpt5}$ | 0.21 | 60.3 | 0.20 | 0.29 | **64.6** | 0.30 | **74.5** | 0.53 | 54.2 | 67.0 | 59.9 | **36.1** |
| | `Gptoss-120B` | | | | | | | | | | | | |
| 11 | Fewshot | 0.28 | 65.2 | 0.29 | 0.33 | 65.4 | 0.33 | 70.9 | 0.53 | 44.3 | **77.3** | 56.4 | 37.1 |
| 12 | FineSURE | 0.18 | 59.3 | 0.18 | 0.30 | 65.3 | 0.30 | 73.1 | 0.50 | 52.3 | 63.2 | 57.3 | 35.9 |
| 13 | VG-prompt | 0.10 | 54.4 | 0.10 | **0.34** | **69.2** | **0.35** | 71.5 | 0.53 | 46.0 | 75.1 | 57.1 | 37.1 |
| 14 | **DESCRIBE** | 0.15 | 57.3 | 0.16 | 0.30 | 67.8 | 0.31 | 74.6* | 0.53 | 54.9* | 65.6 | 59.8* | **35.4** |
| 15 | **DESCRIBE**$^{gpt5}$ | **0.36** | **68.5***| **0.36*** | 0.30 | 64.8 | 0.30 | **75.1*** | **0.54** | **55.8*** | 66.5 | **60.7*** | **35.4** |
| | `Gemini-3-flash` | | | | | | | | | | | | |
| 16 | Fewshot | 0.25 | 63.8 | 0.27 | 0.34 | 64.9 | 0.36 | 72.6 | 0.54 | 48.3 | 74.9 | 58.8 | 36.6 |
| 17 | FineSURE | **0.40** | **71.6** | **0.42** | **0.47** | **70.1** | **0.49** | 71.8 | 0.53 | 47.0 | 74.9 | 57.5 | 38.1 |
| 18 | VG-prompt | 0.24 | 64.2 | 0.30 | 0.32 | 61.8 | 0.37 | 73.9 | 0.60 | 49.5 | 85.4 | 62.7 | **35.6** |
| 19 | **DESCRIBE** | 0.31 | 67.4 | 0.35 | 0.31 | 62.2 | 0.34 | 75.8* | 0.63 | 53.2* | 86.9 | 66.0* | 36.1 |
| 20 | **DESCRIBE**$^{gpt5}$ | 0.32* | 68.6* | 0.38* | 0.38 | 65.0 | 0.41 | **76.8*** | **0.65*** | **55.0*** | **87.8** | **67.6*** | 35.8 |
| | `GPT-5` | | | | | | | | | | | | |
| 21 | Fewshot | 0.25 | 62.4 | 0.26 | 0.40 | 72.0 | 0.40 | 81.3 | 0.66 | 66.7 | 76.1 | 71.0 | 31.3 |
| 22 | FineSURE | 0.35 | 66.4 | 0.37 | 0.42 | **75.0** | **0.43** | 81.6 | 0.66 | 67.8 | 74.3 | 70.9 | 32.0 |
| 23 | VG-prompt | 0.34 | 66.9 | 0.34 | **0.43** | 72.9 | **0.43** | 82.5 | 0.64 | 71.0 | 69.8 | 70.4 | 31.6 |
| 24 | **DESCRIBE** | **0.39** | **69.4** | **0.39** | 0.41 | 71.4 | 0.42 | **84.9*** | **0.73*** | **73.2** | **81.5*** | **77.1*** | **30.8** |

Table 2: **Performance of faithfulness evaluation methods on long, coherent summaries from StorySumm and VeriGray.** For StorySumm, both summary- and sentence-level metrics are reported; for VeriGray only sentence-level metrics, following the original protocol. Metrics include Cohen's Kappa (Coh.-k), balanced accuracy (BAcc.), Matthew's Correlation Coefficient (MCC), hallucination precision (Hal. P.), recall (Hal. R.), F1, and ranking loss ($L_{rank}$). The bottom block shows results across different backbone LMs. Higher is better for all metrics except $L_{rank}$. **Bold** = best results within the block. Underline = not significantly different from best ($p > 0.05$, paired bootstrap resampling test; Appendix F). * means our model is significantly better than the VG-prompt baseline on the classification metrics.

**Impact of discourse-guided approaches across model scales.** StructScore significantly outperforms the MiniCheck baseline (row 4 vs. row 1) on both datasets, while FDSpotter underperforms fine-tuned fact-checkers on VeriGray. For prompt-based methods, performance generally improves as the underlying LLMs grow larger and stronger on general benchmarks (from `Qwen-3` to the strongest `GPT-5`). For VeriGray, our discourse-inspired approach consistently outperforms the best baseline, *VG-prompt*, by ∼2-point gains in balanced accuracy, with hallucination F1 gains increasing with model scale (+2.7 F1 for `gpt-oss`, rows 13–14; +3.3 F1 for `Gemini-3`, rows 18–19; +6.7 F1 for `GPT-5`, rows 23–24). Moreover, replacing claims generated by smaller models with GPT-5-generated claims yields comparable or better results across backbones (rows 14-15, 19-20), demonstrating the benefit of high-quality structured decomposition and the framework's modularity.

For StorySumm, results are mixed. Describe improves over *VG-prompt* at the summary level, but the baseline *FineSURE* sometimes outperforms it at the sentence-level. This likely stems from a mismatch between Describe's discourse-level distinctions and the dataset's binary annotation, which marks sentences unfaithful only under full annotator agreement, masking subtle relational errors that our method detects. A recent work (Zhong et al., 2026) audited StorySumm and found that the unanimity annotation rule suppressed such errors that our method is designed to detect. After correcting the annotation inconsistencies, Describe outperforms other baselines significantly on both datasets. In contrast, VeriGray's seven-category faithfulness taxonomy with expert rationales better captures these distinctions, explaining the consistent gains observed and supporting prior findings on subjectivity in faithfulness evaluation (Subbiah et al., 2025). Overall, discourse-guided methods facilitate finer-grained error analysis.

### 6.2 Impact of Decomposition and Inference

**Does explicitly modeling discourse in decomposition help reduce decomposition errors?** We conduct a similar analysis to §3 on claims generated by DESCRIBE. Compared to FABLES, DESCRIBE decomposed claims are found to contain fewer structurally harmful errors (13.3%). On the combined corpora, the count of `Alteration of meaning` and `Missing Logical Relationships` decreased from 70 to 57, and 32 to 11, respectively. DESCRIBE also maintains comparable ambiguity error rates (7.7% vs. 7.8% for FABLES), indicating that explicitly incorporating discourse relations with the help of discourse parsers helps preserve the contextual integrity and inter-claim dependencies (full tables in Appendix G.1).

**How do different modules contribute to faithfulness assessment?** Table 3 reports the effects of discourse-guided decomposition and structured inference on VERIGRAY via ablations. Rows 1–5 isolate the impact of claim quality by fixing structured inference fixed while varying claim decomposition. We compare DESCRIBE without predicted relations, two discourse-inspired decomposition methods (Fables and VeriFact, §3), and original sentences. DESCRIBE performs best: removing discourse relation modeling (ROW 2) reduces balanced accuracy by 1.7 points, while alternative decomposition methods (ROWS 3-4) and original sentences (ROW 5) further degrades performance.

Rows 6–7 test transfer to simpler prompting setup. We modify the VG-prompt baseline prompt with DESCRIBE-generated claims and relations, instructing the model to leverage cross-claim discourse relations during evaluation. With the VG-prompt inference baseline, our discourse-guided claims (ROW 6) still outperform original sentences (ROW 8) and a baseline that instructs the LLM to perform discourse analysis before evaluation (ROW 7), confirming that higher quality decomposition adds value independent of the inference strategy. Full prompts are provided in Appendix G.2.

**Does discourse modeling help with complex sentence evaluation?** To investigate whether DESCRIBE's gains depend on the discourse complexity of summary sentences, we stratify VERIGRAY by connective presence using a retrained connective predictor (Appendix G.3). Sentences are grouped into those without discourse connectives and those containing connectives such as *"and", "because", and "but"*.[5] Figure 4 shows DESCRIBE's improvement over three baselines on each group using GPT-5. Gains in BAcc. and MCC are consistently larger for sentences with connectives (e.g., MCC gain over VG-prompt increases from +5.5 to +8.1), indicating that discourse-guided decomposition better exploits explicit relational structure than surface-level semantic matching baselines.

| ID | Claim | BAcc | MCC | Hal. R.↑ | Hal. P↑ | Hal. F1↑ | $L_{rank}$ ↓ |
|---|---|---|---|---|---|---|---|
| *Fixed structured inference, varying decomposition* | | | | | | | |
| 1 | **DESCRIBE** | 84.9 | 0.73 | 73.2 | 81.5 | 77.1 | 30.8 |
| 2 | w/o rel. | 83.2* | 0.70* | 69.8* | 79.7 | 74.4* | 31.6 |
| 3 | Fables | 82.6* | 0.69* | 68.7* | 80.0 | 73.9* | 31.8 |
| 4 | Verifact | 82.0* | 0.67* | 68.2* | 76.2* | 71.9* | 31.9 |
| 5 | Original Sent | 81.1* | 0.67* | 66.0* | 77.8* | 71.4* | 32.3 |
| *Fixed VG-prompt baseline inference* | | | | | | | |
| 6 | DESCRIBE | 84.7* | 0.69* | 75.2 | 72.4* | 73.8* | 30.3 |
| 7 | disco-aware | 83.3* | 0.66* | 72.5 | 70.7* | 71.6* | 31.0 |
| 8 | Origin sent | 82.5* | 0.64* | 71.0* | 69.8* | 70.4* | 31.6 |

Table 3: **Ablation study analyzing the effects of decomposition and inference strategies on VERIGRAY, using GPT-5 backbone.** * $p < 0.05$ compared to full model (paired bootstrap).

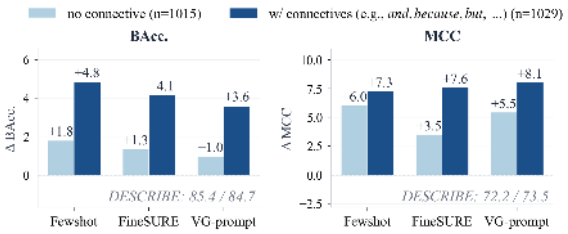



Figure 4: **DESCRIBE's advantage over GPT-5 baselines on stratified sentence groups in VERIGRAY.** Sentences are split by the presence of predicted discourse connectives. Bars show Δ absolute differences; DESCRIBE's absolute scores are in *italics*.

### 6.3 Do Models Provide Trustworthy Rationales?

Recent benchmarks started collecting human rationales alongside faithfulness labels, but these are rarely incorporated into the evaluation process. Seo et al. (2025) and Si et al. (2026) propose prompting LLMs to assess the *completeness*, *informativeness*, and internal *logical consistency* dimensions of generated rationales, yet neither were able to leverage the human-written rationale. We build on these efforts with three assessment dimensions: **Human Rationale Alignment** (H.Align) measures whether model reasoning captures the same justifications as human annotators. **Discourse Awareness** (Disc.)

[5] This grouping may not fully separate discourse-heavy sentences from simpler statements, as implicit discourse relations can still exist in the first group.

| System | $n_1$ | Compl. | $n_2$ | H.Align | Disc. (↑ / ↓ / –) |
|---|---|---|---|---|---|
| STORYSUMM | | | | | |
| DESCRIBE | 496 | 92.9 | 42 | 95.2 | 92.9 / 2.3 / 4.8 |
| VG-prompt | 487 | 80.5 | 53 | 79.2 | 92.5 / 5.6 / 1.9 |
| FineSURE | 476 | 59.0 | 58 | 77.6 | 79.3 / 13.8 / 6.9 |
| FaithLens | 481 | 86.9 | 32 | 18.8 | 68.8 / 25.0 / 6.2 |
| VERIGRAY | | | | | |
| DESCRIBE | 1741 | 95.2 | 343 | 92.7 | 92.1 / 1.2 / 6.7 |
| VG-prompt | 1746 | 88.5 | 357 | 90.5 | 82.4 / 5.9 / 11.8 |
| FineSURE | 1734 | 70.4 | 330 | 85.2 | 74.8 / 11.8 / 13.3 |
| FaithLens | 1581 | 92.4 | 193 | 72.0 | 58.5 / 22.8 / 18.7 |

Table 4: **Rationale evaluation (%) on correctly predicted instances.** $n_1$: instances where $\hat{y} = y_{\text{gold}}$; $n_2$: subset of $n_1$ with human rationales. Evaluation covers completeness (Compl.), human rationale alignment (H.Align), and discourse awareness (Disc.), reported as proportions of correct (↑) / incorrect (↓) / N/A (–).

evaluates whether systems correctly identify and validate faithfulness-relevant discourse relations. **Completeness** (Compl.) measures the thoroughness of the evaluation protocol's evidence inspection, though we note this dimension is influenced by each system's output design.

**Automatic Evaluation.** Table 4 shows the rationale evaluation results using `Gemini-3-Flash` as the evaluation backbone.[6] Following Si et al. (2026), we evaluate only correctly predicted instances to isolate rationale quality from prediction errors. DESCRIBE achieves the highest Completeness on both datasets, outperforming the next best system by at least 2.8 points, suggesting that discourse-guided decomposition improves claim-level verification and reduces missing evidence in rationales. Due to the limited availability of human rationales for STORYSUMM, we focus our analysis of Human Rationale Alignment and Discourse Awareness on VERIGRAY. DESCRIBE achieves the highest Human Rationale Alignment score, indicating closer alignment with human reasoning. The Discourse Awareness breakdown further supports this finding: DESCRIBE correctly identifies discourse information in over 94% cases, while the other systems show substantially higher error rates.

Analysis of incorrectly predicted instances reveals that DESCRIBE still produces substantially more reliable rationales, with higher correctness rates in both Human Rationale Alignment (39.8% vs. 30.9%) and Discourse Awareness (83.7% vs. 64.7%) than VG-prompt. This suggests that even when predictions are incorrect, DESCRIBE's reasoning remains better aligned with human judgment and discourse structure, highlighting a gap between label accuracy and rationale quality further explored in an error analysis presented in §7.

**Manual Evaluation.** We conduct a human evaluation of model-generated rationales (Table 2, ROWS 20–22) using the same protocol. We sample 50 instances per model, yielding 150 total annotations—exceeding prior studies (Wadhwa et al., 2024; Si et al., 2026), which typically evaluate 50–120 annotations. Using the adjudicated annotations between two annotators and the first author as gold labels, the automatic evaluation of the accuracy/macro-F1 of Completeness, Human Rationale Alignment, and Discourse Awareness are 0.91/0.82, 0.96/0.93, and 0.92/0.84, respectively (Appendix H.4).

## 7 Error Analysis

We sample 100 instances where DESCRIBE predicts incorrect labels and identify three main error categories: (1) inconsistent rationale and label (16.7%), where the model surfaces correct evidence but assigns the wrong category label; (2) over-interpretation (25.0%), where implied information is treated as explicitly entailed from the original document; and (3) missing fine-grained details (36.0%), where the model verifies surface-level content but overlooks key relational structures such as temporal framing or comparison. The latter two categories match the two most frequent errors in Seo et al. (2025). The remaining 22.3% involve less representative errors.

Our qualitative inspection reveals three recurring error types. Firstly, DESCRIBE can generate correct rationales yet still assign an incorrect faithfulness label, revealing a mismatch between reasoning quality and final prediction. Secondly, the model may overinterpret the source by justifying inferred claims not explicitly supported by the document. Thirdly, despite discourse-aware reasoning, it can still miss subtle semantic details that change the original meaning.[7] Together, these findings suggest that while DESCRIBE improves evidence identification and discourse relation verification, translating nuanced reasoning into accurate categorical faithfulness labels remains challenging.

[6] We chose it to avoid the self-favoring effect, as the rationales are generated by GPT-5. Assessments using GPT-5-mini show similar patterns (Appendix H.2).

[7] See concrete examples in Table 14.

## 8 Conclusion

We present DESCRIBE, a discourse-informed framework for evaluating the faithfulness of long-form abstractive summaries. Our analysis of existing discourse-guided claim decomposition methods shows that on average over 40% of decomposed outputs contain errors that disrupt the original summary's semantic context and discourse structure. To address this, we introduce a discourse-informed claim extraction approach based on RST, that preserves intra- and inter-sentential relations, paired with a structured inference procedure for hierarchical validation of claims and relations.

Experiments on two benchmarks demonstrate consistent improvements over fine-tuned fact-checkers and prompt-based baselines. Beyond prediction accuracy, DESCRIBE produces more complete rationales with stronger human alignment and discourse awareness. Error analysis further highlights the gap between rationale quality and label prediction. More broadly, our findings indicate that discourse-aware representations provide a promising direction for improving both the reliability and interpretability of faithfulness evaluation in long-form summarization.

## Limitations

As a result of using the STORYSUMM and VERIGRAY benchmarks, our findings are limited to English-language summaries, excluding multilingual scenarios. Generalization to other domains and languages can be a promising future direction. Our framework depends on discourse parsing quality, and errors in RST tree extraction may propagate to claim decomposition and then to faithfulness judgments. We retrain a discourse parser, similar to prior work (Liu et al., 2023a,c), using the largest available RST dataset, which covered diverse domains of data.[8]

Ablations on direct prompting LLMs with demonstration examples show degraded downstream task performance (Appendix G.4), and recent work (Liu et al., 2026) has also demonstrated that a zero-shot parser underperformed a trained discourse parser. Computation-wise, DESCRIBE incurs additional cost from the claim decomposition step; we conduct a detailed cost analysis, showing that the effective overhead remains modest with prompt caching ($0.12 vs. $0.05 per summary using GPT-5) in Appendix E.7. The DMRST parser used to predict the parsed RST tree can be run on both CPU and GPU, and the inference speed is fast (the full VERIGRAY can be processed in a few minutes), thus introducing minimal overhead. We also acknowledge that faithfulness judgments involve inherent subjectivity, particularly at fine-grained levels, and our framework may inherit the backbone LLMs' biases in making predictions.

We also encourage future annotations of faithfulness errors to extend beyond sentence-level granularity. While prior work has annotated text spans with fixed error categories (Cao and Wang, 2021), our results suggest that relation-level errors across claims constitute a meaningful and distinct error dimension; however, our quantitative evaluation remains bounded by sentence-level labels, as no existing benchmark provides sub-sentential relation annotations. Exploring the practical utility of rationale qualities would be an important direction.

Lastly, we acknowledge that potential noise and subjectivity in the used datasets may affect the evaluation results. Our analysis of STORYSUMM identified two areas for improvement: first, the inclusion of 19 sentences beyond the original task instructions and second, the dataset's reliance on a unanimous consensus rule for labeling unfaithfulness. While designed for high precision, this strict agreement threshold may inadvertently overlook nuanced errors, potentially resulting in false-positive faithful labels. To maintain comparability with existing benchmarks, we report our results following the original paper's protocols, noting that these inherited annotation characteristics may persist in the final metrics.

## Ethical Considerations

Benchmarking datasets used in this paper are both publicly available for research purposes. We acknowledge the potential for bias in human evaluation, as both in-house annotators are based in the U.S. and hold a master's degree with experience in natural language processing tasks. Given the narrative genre of STORYSUMM, one evaluator reported a feeling of discomfort in trial runs due to the horror content, and we opted to mainly analyze VERIGRAY for manual assessments. No personal information is collected from the annotators.

Our framework relies on LLMs for both claim decomposition and structured inference, which may reflect biases present in these LLMs' training data. We emphasize that DESCRIBE is intended as a diag-

[8] https://gucorpling.org/gum/

nostic research tool for evaluating summary faithfulness and should not be deployed as a standalone factual checker in high-stakes domains (e.g., medical or legal) without human expert oversight.

We acknowledge the use of Grammarly and AI assistants (ChatGPT and Claude) for correcting sentences that are less fluent, but not for generating or drafting new content.

## 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 for their valuable suggestions, which helped improve this work. We also thank Zhehan Tiffany Zhu and Arun Balajiee Lekshmi Narayanan for their contributions to the human evaluation.

## References


Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes. 2023. The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada. The Association for Computational Linguistics.

Chloé Braud, Amir Zeldes, Chuyuan Li, Yang Janet Liu, and Philippe Muller. 2025. The DISRPT 2025 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 4th Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2025)*, pages 1–20, Suzhou, China. Association for Computational Linguistics.

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

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, and 81 others. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

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

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Davide Chicco, Matthijs J. Warrens, and Giuseppe Jurman. 2021. The matthews correlation coefficient (mcc) is more informative than cohen's kappa and brier score in binary classification assessment. *IEEE Access*, 9:78368–78381.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

Rosario Delgado and Xavier-Andoni Tibau. 2019. Why cohen's kappa should be avoided as performance measure in classification. *PLOS ONE*, 14(9):1–26.

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*.

Pingjun Hong, Beiduo Chen, Siyao Peng, Marie-Catherine de Marneffe, and Barbara Plank. 2025. LiTEx: A linguistic taxonomy of explanations for understanding within-label variation in natural language inference. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34065–34085, Suzhou, China. Association for Computational Linguistics.

Qisheng Hu, Quanyu Long, and Wenya Wang. 2025. Decomposition dilemmas: Does claim decomposition boost or burden fact-checking performance? 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 6313–6336, Albuquerque, New Mexico. Association for Computational Linguistics.

Nazanin Jafari, James Allan, and Mohit Iyyer. 2026. Beyond precision: Importance-aware recall for factuality evaluation in long-form llm generation. *Preprint*, arXiv:2604.03141.

Yeonseok Jeong, Minsoo Kim, Seung-won Hwang, and Byung-Hak Kim. 2025. Agent-as-judge for factual summarization of long narratives. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23591–23608, Suzhou, China. Association for Computational Linguistics.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. FABLES: Evaluating faithfulness and content selection in book-length summarization. In *First Conference on Language Modeling*.

Wei-Jen Ko and Junyi Jessy Li. 2020. Assessing discourse relations in language generation from GPT-2. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 52–59, Dublin, Ireland. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. 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, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Dawon Lee, Hyuckchul Jung, and Yong Suk Choi. 2025. Mind the link: Discourse link-aware hallucination detection in summarization. *Applied Sciences*, 15(19).

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.

Dongqi Liu, Hang Ding, Qiming Feng, Xurong Xie, Zhucun Xue, Chengjie Wang, Jian Li, Jiangning Zhang, and Yabiao Wang. 2026. Disco-RAG: Discourse-aware retrieval-augmented generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4106–4136, San Diego, California, United States. Association for Computational Linguistics.

Dongqi Liu, Yifan Wang, and Vera Demberg. 2023a. Incorporating distributions of discourse structure for long document abstractive summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5574–5590, Toronto, Canada. Association for Computational Linguistics.

Wei Liu, Yi Fan, and Michael Strube. 2023b. HITS at DISRPT 2023: Discourse segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 43–49, Toronto, Canada. The Association for Computational Linguistics.

Xin Liu, Lechen Zhang, Sheza Munir, Yiyang Gu, and Lu Wang. 2025. VeriFact: Enhancing long-form factuality evaluation with refined fact extraction and reference facts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17908–17925, Suzhou, China. Association for Computational Linguistics.

Yang Janet Liu, Tatsuya Aoyama, and Amir Zeldes. 2023c. What's hard in English RST parsing? predictive models for error analysis. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–42, Prague, Czechia. Association for Computational Linguistics.

Yang Janet Liu and Amir Zeldes. 2023. Why can't discourse parsing generalize? a thorough investigation of the impact of data diversity. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3112–3130, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. DMRST: A joint framework for document-level multilingual RST discourse segmentation and parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 154–164, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723.

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.

B.W. Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Eleni Metheniti, Chloé Braud, Philippe Muller, and Laura Rivière. 2023. DisCut and DiscReT: MELODI at DISRPT 2023. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 29–42, Toronto, Canada. The Association for Computational Linguistics.

Yisong Miao, Hongfu Liu, Wenqiang Lei, Nancy Chen, and Min-Yen Kan. 2024. Discursive socratic questioning: Evaluating the faithfulness of language models' understanding of discourse relations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6277–6295, Bangkok, Thailand. Association for Computational Linguistics.

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

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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.

Siyao Peng, Yang Janet Liu, and Amir Zeldes. 2022. GCDT: A Chinese RST treebank for multigenre and multilingual discourse parsing. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 382–391, Online only. Association for Computational Linguistics.

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. Penn Discourse Treebank Version 3.0.

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

Shuzheng Si, Qingyi Wang, Haozhe Zhao, Yuzhuo Bai, Guanqiao Chen, Kangyang Luo, Gang Chen, Fanchao Qi, Minjia Zhang, Baobao Chang, and Maosong Sun. 2026. FaithLens: Detecting and explaining faithfulness hallucination. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 14068–14099, San Diego, California, United States. Association for Computational Linguistics.

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, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Manya Wadhwa, Xinyu Zhao, Junyi Jessy Li, and Greg Durrett. 2024. Learning to refine with fine-grained natural language feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12281–12308, Miami, Florida, USA. Association for Computational Linguistics.

David Wan, Koustuv Sinha, Srini Iyer, Asli Celikyilmaz, Mohit Bansal, and Ramakanth Pasunuru. 2024. ACUEval: Fine-grained hallucination evaluation and correction for abstractive summarization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10036–10056, Bangkok, Thailand. Association for Computational Linguistics.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Zixia Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V Le. 2024. Long-form factuality in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao,

Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Joonho Yang, Seunghyun Yoon, ByeongJeong Kim, and Hwanhee Lee. 2024. FIZZ: Factual inconsistency detection by zoom-in summary and zoom-out document. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Miami, Florida, USA. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Huajian Zhang, Yumo Xu, and Laura Perez-Beltrachini. 2024. Fine-grained natural language inference based faithfulness evaluation for diverse summarisation tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1701–1722, St. Julian's, Malta. 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.

Shiyue Zhang, David Wan, and Mohit Bansal. 2023. Extractive is not faithful: An investigation of broad unfaithfulness problems in extractive summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2153–2174, Toronto, Canada. 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.

Yang Zhong, Yang Janet Liu, and Diane Litman. 2026. Revisiting faithfulness annotations for long-form summaries. In *Proceedings of the 20th Linguistic Annotation Workshop (LAW XX)*, pages 158–172, San Diego, California, USA. Association for Computational Linguistics.

## A Extended Discussion of Related Work

In addition to the decompose-then-verify paradigm, researchers explore other approaches to assess long summaries. Zhong and Litman (2025a) use discourse structure to readjust sentence-level NLI scores rather than verifying segmented decomposed claims. NarrativeFactScore (Jeong et al., 2025) evaluates factuality through an agentic framework built on character knowledge graphs. We also note work that extends the paradigm by encouraging recall-aware evaluation, which scores coverage of salient source facts alongside precision (Jafari et al., 2026).

**Comparison to other discourse-inspired approaches.** We discuss the differences between our work and prior discourse-guided frameworks. Compared to StructScore (Zhong and Litman, 2025a), both works use RST and show that improvements are concentrated in discourse-complex sentences. The key methodological distinction is in how discourse information is used. StructScore uses tree features to weight and aggregate sentence-level scores. DESCRIBE uses discourse structure and relational information to verify claim support and relation fidelity against the source. Secondly, StructScore focuses on score-aggregation error, whereas DESCRIBE's focus on relation fabrication and nuanced errors, which reweighting cannot detect. On output and backbone differences, StructScore emits a reweighted document-level consistency score and requires a continuous scalar scorer. DESCRIBE provides fine-grained labels with rationales and runs on label-only backbones that StructScore's weighting cannot consume.

Another baseline relevant to our work is from Lee et al. (2025). Their work models only three AMR discourse links (cause, purpose, and time), whereas DESCRIBE reasons over the full RST relation inventory, including contrast, attribution, elaboration, explanation, and other implicit relations. Meanwhile, the Discourse Link-Aware Content Unit (DL-ACU) framework captures pairwise links between content units and serves as a preprocessing module for downstream NLI-based detectors. DESCRIBE performs document-level, relation-aware verification as a standalone zero-shot evaluator, producing fine-grained error labels and interpretable natural-language rationales.

## B Claim Decomposition Methods

In this section, we cover the implementation details for the two reported discourse-aware claim decomposition approaches used in §3. We acknowledge that there exist more diverse prompt types for deriving claims from a summary (Scirè et al., 2024; Yang et al., 2024; Wan et al., 2024). However, after manually inspecting their prompts, we find that many rely on simple one-sentence breakdown examples or biography-domain demonstrations, which could be less useful for handling the long-text decomposition scenario. We opt not to include generic decompositions in the claim-checking domain (i.e., FactScore (Min et al., 2023) and SAFE (Wei et al., 2024)), as they have been found to suffer from incomplete and missing facts in the VeriFact paper (Liu et al., 2025), which is one of our baselines. Therefore, we focus on more recent approaches that leverage discourse information in the decomposition process.

### B.1 Decomposition Baselines

**Fables** (Kim et al., 2024). We use a slightly modified version of the claim-extraction prompt provided by the original paper (Kim et al., 2024). The original version is tailored for extracting claims from an AI-generated book summary, using specific phrases such as "a given summary of a book". While the prompt is relatively short, it was found to outperform the baseline sentence-level assessment in StorySumm (Subbiah et al., 2024), especially given the literal requirements on preserving the "temporal, location and causal context" in the prompt. The full prompt is included in Figure 5. In our experiments, we prompt GPT-5 to obtain the corresponding claim lists.

**VeriFact** (Liu et al., 2025) is a complicated pipeline work to generate and refine the list of facts with the goal of recovering incomplete and missing facts with the help of discourse-inspired exploration. It is a multi-stage framework that applies claim decomposition, de-contextualization, detection of incomplete and missing facts, and text refinement strategies to tackle the challenges in preserving the coverage and quality of the claims. We follow their official codebase `https://github.com/launchnlp/VeriFact` to produce claim lists for all summaries in our test suite. Different from the paper description, the open-sourced version leverages Llama-3.3-70B-Instruct for claim extraction and later stages, such as LLM-Judge annota-

**Prompt for FABLES Decomposition**

You are trying to verify the faithfulness of statements made in a given summary of a document against the actual text of the document. To do so, you first need to break the summary into a set of "atomic claims", each of which will then be passed to a human who will read the original doc and verify if the claim is true or not. Each atomic claim must be fully understandable without any other context from the summary (e.g., all entities must be referred to by name, not pronoun), and they must be situated within relevant temporal, location, and causal context whenever possible. Try to keep each atomic claim to a maximum of 2 sentences. Each atomic claim is separated with '- '.

**Summary**: [SUMMARY]
**List of atomic claims**:

Figure 5: Prompt used for FABLES claim extraction.

tion and claim refinement. We acknowledge that this reimplementation might not produce the most optimal results as reported in the original paper, which employs three large language models for voting; however, given the available referenced sources, it provides a reasonable approximation. We utilize the merged claim lists produced by different refinement approaches after deduplication for our analysis.

### B.2 Decomposition Error Assessment

Following Hu et al. (2025), we prompt GPT-5 with their open-source prompts[9] to assess different decomposition lists given the original summary. Table 5 shows examples for each error category in addition to over-decomposition. We acknowledge that the automatic evaluation is not perfect, particularly for the classification of over-decomposition categories, which are of less interest compared to the primary error types. Similar to prior work, we conduct a small-scale human evaluation by sampling 90 decomposition outcomes from all datasets across three methods, revealing a 91% agreement rate (here we mainly focus on error types other than over-decomposition, as it is extremely pervasive on VeriFact and does not account much for the key issues). We additionally conduct a manual check on the assessments of our approach DESCRIBE (§6), since we observe that the evaluator sometimes fails to leverage the relation field, which encodes the core discourse relations between claims.

[9] https://github.com/qishenghu/Decomp_Dilemmas/blob/main/src/third_party/self_diagnosis.py

## C Discourse Parser Retraining

**Dataset Choice.** Following prior work (Peng et al., 2022; Liu et al., 2023c), we retrain the DMRST parser (Liu et al., 2021) from scratch by leveraging the GUM v11 (Zeldes, 2017).[10] We leverage the original train/dev/test splits, and Table 6 provides an overview of the data. Compared to RST-DT (Carlson et al., 2001), which lacks an established separate dev set and contains only news domain data, GUM covers diverse text genres ranging from academic writing to podcasts. We utilize the Level-1 relations, which cover 15 relation classes.

**Implementation Details.** We initialize the model with the RoBERTa-base checkpoint.[11] The learning rate is set as 5e-5, and we keep other parameters the same as the official code base. Following Liu and Zeldes (2023), we run the training five times with different seeds, keeping the checkpoint that obtains the highest validation performance. Following prior work, we report on three metrics: Span (whether subtrees span the right EDUs), Nuclearity (whether edges point the right way), and Relation (whether labels are correct) using the stringent original Praseval procedure on binary trees, with the average scores of 71.9, 50.4, and 58.5 for Span, Relation, and Nuclearity, respectively. For inference on raw text, we use the checkpoint with the highest relation scores.

## D Implementation Details of DESCRIBE

### D.1 Final Fine-grained Labels

We adopt the faithfulness taxonomy introduced by Ding et al. (2025), which offers a fine-grained assessment to handle cases that are ambiguous or require external sources' knowledge. We borrow the illustrative decision tree figure from the original paper in Figure 6. Table 7 presents the definition and examples of each label.

### D.2 RST Tree Prediction

To predict the RST tree of plan-text summaries in the evaluation dataset, we adopt a two-stage setup to guarantee the quality of the parsed tree. While the original DMRST (Liu et al., 2021) offered a joint discourse EDU segmentation and discourse-tree prediction mode to process raw texts, it was

[10] We utilize the released dataset in https://github.com/amir-zeldes/gum/tree/master/rst

[11] https://huggingface.co/FacebookAI/roberta-base

| Error Type | Summary | Decomposition |
|---|---|---|
| Omission Logical Relationships | ... Hardy, ... is also an expert on the medieval longbow. He plans to use the auction to spread joy and pleasure to new owners, as downsizing brings him sadness. | • Robert Hardy is an expert on the medieval longbow.<br>• Robert Hardy plans to use the auction to spread joy and pleasure to new owners.<br>• Downsizing his home brings Robert Hardy sadness.<br>**Justification**:<br>The causal relationship signaled by **"as"** is diminished. |
| Omission Core Claim | McGregor claims he is more motivated to win the title than Aldo ... The two fighters have a tense relationship highlighted by an incident during a promotion tour in Dublin ... | • Conor McGregor claims he is more motivated to win the feather-weight title than Jose Aldo ...<br>• During a promotional tour event in Dublin, ...<br>**Justification**: Omits the explicit characterization that "the two fighters have a tense relationship". |
| Alternation of Original Meaning | Tomorrow finds comfort in his memory of the girl and the way she said his name, forever determined to meet him, to meet Tomorrow. | • ... Tomorrow finds comfort and warmth in his memories of the way the little girl said his name.<br>• Tomorrow remains determined to meet the little girl again.<br>**Justification**:<br>Misattributes the determination. It refers to the girl, not the dog. |
| Ambiguity | Barrow centre-back Anderson, has joined Torquay United ... | • Anderson, a 27-year-old centre-back from Barrow, has joined Torquay United.<br>**Justification**:<br>The claim changes "Barrow centre-back" (indicating Anderson played for Barrow) to "centre-back from Barrow", which can imply place of origin. |

Table 5: **Examples of decomposition errors following the error typology of** Hu et al. (2025). We highlight error attributions within the summary, providing LLM-generated claims alongside system justifications.

| | RST-DT | GUM v11 |
|---|---|---|
| **# of docs** | 385 | 281 |
| *train/dev/test/(OOD)* | 347/-/38/0 | 191/32/32/26 |
| **# of tokens** | 203,352 | 268,208 |
| **# of EDUs** | 21,789 | 35,144 |
| **# of genres** | 1 | 16/(8) |
| **# of relation labels** | 78 | 32 |
| **# of relation classes** | 17 | 15 |

Table 6: Statistics of the English RST Corpora.

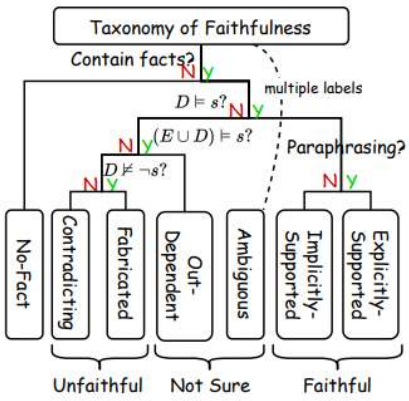



Figure 6: **Decision tree for annotating unfaithfulness.** The Ambiguous label is not assigned by a single run of the decision tree. Instead, it is applied when the procedure yields multiple labels for different interpretations of the same sentence.

found that the segmentation module might not yield optimal performance. To resolve this challenge, we leverage the best-performing EDU segmentation model, DISCUT (Metheniti et al., 2023), in the shared task of DISRPT 2023 (Braud et al., 2023) to extract EDUs from the raw texts, then apply our best-trained model to infer the discourse tree. The segmenting model obtains 94.46 test F1 on `en.rst.gum` split. We use the `.dis` file in LISP format for later-stage claim decomposition. In the following subsections, we describe the prompts employed for our approaches.

### D.3 Discourse-guide Decomposition

We include the complete prompt for our proposed decomposition approach in Figure 16. The first paragraph is adopted from FABLES (Kim et al., 2024) by modifying the original wording related to the book summary. We also append the role of "expert in Computational Linguistics" in the instruction. Additionally, we introduce the RST coarse-grained relations from the GUM dataset (Zeldes, 2017) to guide step 2 and step 3 relation derivation.[12] We include three demonstrative examples consisting of paragraphs extracted from GUM, along with their gold discourse parsing trees and human-verified decomposition JSON files.

### D.4 Structured Inference

We include the complete structured inference instruction in Figure 17 and Figure 18.

[12] https://wiki.gucorpling.org/gum/rst

| Document | Target Sentence | Annotation |
|---|---|---|
| "the black box" of the Su-24 jet was officially opened in Moscow on Friday in front of journalists and diplomats. . . . The Su-24 was shot down by F-16 fighters on 24 November. | The "black box" of the Su-24 jet, shot down on November 24th, was opened in Moscow on Friday. | `Explicitly-Supported`. |
| the city of eugene in the united states has been awarded the rights to host the 2021 iaaf world championships. . . . the 2007 world championships held in osaka, japan was also awarded without a bidding process . . . | This is not the first time a World Championships was awarded without a bidding process, as the 2007 event in Osaka, Japan was also granted in this manner. | `Implicitly-Supported`. **Reason:** The document supports the claim but does not explicitly mention the highlighted text. |
| . . . the ace was nicklaus' first on the hallowed turf of augusta, but the veteran saw his achievement eclipsed by that of camilo villegas on a day which saw five hole-in-ones, levelling the record from 2002. . . . | Despite his achievement, Camilo Villegas stole the show by scoring two holes-in-one, tying the record. | `Fabricated`. **Reason:** The record in the document does not refer to the two ace of Villegas. |
| conor mcgregor is gearing up for a featherweight title challenge against jose aldo on july 11. . . . aldo makes the eighth defence of his belt against the irish fighter in las vegas, but mcgregor claimed last week that the man he is challenging lacks the same motivation as him. | This is McGregor's eighth defense of the title, and the two fighters have a contentious relationship. | `Contradicting`. **Reason:** It's Aldo's 8th defense. |
| the zhangzhou plant was slated for xiamen . . . however, it provoked an angry backlash in 2007 due to pollution concerns and prompted the local government to relocate the factory to its current, more remote location. | This plant had experienced a previous explosion in 2013 and was relocated from Xiamen due to pollution concerns and protests in 2007. | `Ambiguous`. **Reason:** While the sentence is accurate when "in 2007" modifies "protests," the alternative reading—that it specifies the time of relocation—lacks support. |
| at the grand old age of 75 , jack nicklaus is still capable of hitting aces . . . nicklaus became the youngest person to wear a green jacket in 1963 , and collected his sixth in 1986 . he is one of five men to complete the career grand slam , an accolade which favourite rory mcilroy can achieve if he wins his third major in succession. | Jack Nicklaus is a renowned golfer, having won the Masters Tournament six times, including being the youngest person to wear a green jacket in 1963. | `Out-Dependent` **Reason**: It requires external world knowledge (see `https://en.wikipedia.org/wiki/Augusta_National_Golf_Club`) to interpret wearing a green jacket as winning a Masters Tournament championship. |
| Russia has demanded an apology from Turkey and in the meantime has imposed sanctions including a ban on package holidays, which could cost Turkey billions of dollars. | Russia demands an apology from Turkey. Word Count: 150 | `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 7: **Examples of the taxonomy defined in** VERIGRAY (Ding et al., 2025). The key segment that drives the annotation decision is in red. Table credit to the original authors.

# E Additional Implementation Details

## E.1 Specialized Fact-Checker Baselines

**BeSpoke-MC-7B (MiniCheck).** We harnessed the SOTA `Llama-3.1-BeSpoke-MiniCheck-7B` (`BeSpoke-MC-7B`) released by BeSpoke Labs. This work was published as an EMNLP 2024 paper, and the released model has been widely used by the community as a strong fact-check baseline. The model is fine-tuned from `internlm/internlm2_5-7b-chat` (Cai et al., 2024) on the combination of 35K data points following the approach in MiniCheck (Tang et al., 2024). We use the suggested code repo from `https://huggingface.co/bespokelabs/BeSpoke-MiniCheck-7B`.

**ClearCheck.** The paper of Seo et al. (2025) found that small fine-tuned models underperform larger models in instances that require complex reasoning (e.g., HoVer dataset, Jiang et al. 2020). To close this gap, they construct a synthetic multi-hop fact verification dataset through crawling diverse Wikipedia documents and employ LLMs to extract facts and produce question-answer pairs according to these facts. Afterwards, they utilize the `Llama-3.1-405B-Inst` model to generate direct answers and Chain-of-thought (CoT) reasoning traces on 57K ANLI data points and 25.2K synthetic data, and finetune a smaller `Llama-3.1-8B` model on this dataset following a multi-task fashion. We leverage the open-source model checkpoint, ClearCheck-8B and experimented with both CoT and Direct answering models. The released model obtains substantial performance gains on tasks that require more complex reasoning, representing a strong candidate to tackle the task of unfaithfulness detection of LLM-generated summaries. We report CoT in the main paper as it provides rationales in addition to the final prediction. For hyper-parameter setting, we follow the official implementation, setting *max_tokens=4, temperature=0.0, top_p=1.0* for the Direct mode and *temperature=0.1* for the CoT setting.

**FaithLens** (Si et al., 2026) is the latest fact-checker that both improves the correctness of predictions and provides explanations. We leveraged

their open-source 8B Llama-3.1 model, which is fine-tuned and RL-ed on synthesized samples with explanations. The model is further strengthened through a rule-based reinforcement learning stage RL to improve both the prediction correctness and the explanation quality. We follow the usage guidelines from https://github.com/S1s-Z/FaithLens/tree/master, with the checkpoint downloaded from https://huggingface.co/ssz1111/FaithLens.

For all models, we adopt the configurations suggested by the original authors, setting the temperature to `0.0`, `0.1`, and `0.6` for MiniCheck, ClearCheck, and FaithLens, respectively.

### E.2 Discourse-guided Methods

**StructScore.** We used the official code repo[13] and computed the reweighted sentence-level score based on MiniCheck predictions.

**FDSpotter.** We followed the authors' released codebase.[14] As the original model, GPT-4o, for fact extraction is no longer accessible, we utilize the GPT-5.6-luna model with reasoning effort at xhigh. For final claim scoring, we loaded the official `Inria-CEDAR/FDSpotter-DeBERTa-V3-Large` checkpoint and aggregated using the arithmetic mean across candidates.

### E.3 LLM Details

**LLM Choices.** We pick two open-source LLMs, `Qwen3-30B-A3B-Thinking-2507` (Yang et al., 2025)[15] and `gpt-oss-120B` (OpenAI, 2025)[16], hosted with vLLM (Kwon et al., 2023) v0.10.0. We also experiment with `GPT-5`[17] and `Gemini-3-Flash`[18] by leveraging their official APIs. We set the temperature to `0.6`, `top_p` at `0.95`, `top_K` at `20` for Qwen3, and temperature at 1 for GPT-oss, following the official suggested setting. For GPT-5, we set the reasoning effort to medium due to budget constraints, and reasoning effort to high for `Gemini-3-Flash`.

### E.4 Evaluation Metrics Details

In this section, we provide details of evaluation metrics included in Section 5.

**Balanced Accuracy** measures the average of sensitivity (true positive rate) and specificity (true negative rate), providing an evaluation that is robust to class imbalance. Formally, it is defined as:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right), \tag{1}$$

where $TP$, $TN$, $FP$, and $FN$ denote the number of true positives, true negatives, false positives, and false negatives, respectively.

**Matthews Correlation Coefficient** (MCC) evaluates the agreement between predicted and gold labels by considering all four entries of the confusion matrix. It is defined as:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{2}$$

MCC ranges from $-1$ to 1, where 1 indicates perfect prediction, 0 indicates random-level performance, and $-1$ indicates complete disagreement. It is found to be more reliable for binary classification tasks (Delgado and Tibau, 2019; Chicco et al., 2021).

**Cohen's Kappa** ($\kappa$) measures the agreement between predicted and gold labels while accounting for agreement occurring by chance. It is defined as:

$$p_o = \frac{TP+TN}{TP+TN+FP+FN}, \tag{3}$$

where $p_o$ is the observed agreement:

$$p_o = \frac{TP+TN}{TP+TN+FP+FN}, \tag{4}$$

$$p_e = \frac{(TP+FP)(TP+FN)+(TN+FP)(TN+FN)}{(TP+TN+FP+FN)^2} \tag{5}$$

The value of $\kappa$ ranges from $-1$ to 1, where 1 indicates perfect agreement, 0 indicates chance-level agreement, and negative values indicate agreement worse than chance.

**Verigray Metrics.** VeriGray (Ding et al., 2025) introduces two protocols. The first protocol removes 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 and the

[13] https://github.com/cs329yangzhong/discourse-driven-summary-factuality-evaluation

[14] https://github.com/guihuzhang/FDSpotter

[15] https://huggingface.co/Qwen/Qwen3-30B-A3B-Thinking-2507

[16] https://huggingface.co/openai/gpt-oss-120b

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

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

Matthews correlation coefficient as introduced above. We further compute the precision, recall, and F1 for the unfaithful class, following the setup in the original paper. Predictions are aligned accordingly: for specialized fact-checkers, we map the 1 / 0 predictions to Faithful and Unfaithful classes (MiniCheck, FaithLens). For ClearkCheck, we map [Attributable] to faithful and the remaining two categories ([Contradictory] and [Not Attributable]) as unfaithful. For LLMs, we map FineSURE's "no error" to faithful and other types of errors to unfaithful. The few-shot model uses the same taxonomy as ClearCheck, so we adopt the same mapping. For the remaining methods with fine-grained predictions that follow the VERIGRAY taxonomy, specifically VG-prompt and DESCRIBE, classes with a faithfulness degree no less than a threshold (defined later) are considered Faithful, with the rest as Unfaithful. Here, the *order of faithfulness degree* is defined as:

$$\begin{aligned}\text{Contradicting} < \text{Fabricated} < \text{Ambiguous} \\ < \text{No-Fact} < \text{Out-Dependent} \\ < \text{Implicitly-Supported} < \text{Explicitly-Supported}\end{aligned} \tag{6}$$

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

The second protocol evaluates the ranking quality of the hallucination detector. Here, we cite the original paper to keep all details complete. Consider the full dataset filtering out No-Fact instances only, denoted as $D_{\text{fact}} = \{(x_i, y_i)\}_{i=1}^{N_{\text{fact}}}$, where $y_i$ is the fine-grained label of instance $x_i$. For predictions, instead, the output space is merged into Faithful/Unfaithful as before, assigned with a faithfulness degree order of Unfaithful $<$ Faithful. Let

$$\mathcal{P} := \{(x_i, x_j) \mid y_i > y_j,\ (x_i, y_i), (x_j, y_j) \in D_{\text{fact}}\} \tag{7}$$

Inspired by the ranking expression of AUC (Area Under the Curve, we report the metric **ranking loss** defined in the original paper:

$$\begin{aligned}L_{\text{rank}} := \frac{1}{|\mathcal{P}|} \sum_{(x_i,x_j)\in\mathcal{P}} (\mathbb{I}[f(x_i) < f(x_j)] \\ + \frac{1}{2}\mathbb{I}[f(x_i) = f(x_j)])\end{aligned} \tag{8}$$

where $f$ is the prediction of the hallucination detector. This metric is non-negative and equals zero if and only if $f$ perfectly preserves the faithfulness ranking of $y$.

We use the official evaluation protocol and modified the label mapping to fit other baselines https://github.com/DingQiang2018/verigray/blob/main/utils.py.

### E.5 LLM Prompts

We include the prompt for FineSURE (Song et al., 2024), a modified version of VG-prompt (Ding et al., 2025) in Table 8, which are adopted from the corresponding paper 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.

### E.6 Machine Configuration for Models

We use up to 4 NVIDIA L40 or A100 GPUs, each equipped with 48 GB VRAM and 40 GB VRAM, respectively, for model inferences on our hardware. According to Lambda,[19] a single NVIDIA A6000 (the closest to our setting with 48GB VRAM) GPU costs $0.8 per hour.

### E.7 Inference Cost Using LLMs

Computation-wise, DESCRIBE requires one additional LLM call for discourse-guided claim decomposition, and appending the decomposed claims into the input for structured-inference, resulting in approximately 3.5x token cost compared to single-pass prompting approaches such as VG-prompt (approximately 22K vs. 6.5K tokens per data instance on STORYSUMM using GPT-5). In practice, over 60% of input tokens consist of shared prompt templates eligible for caching, reducing the effective cost to approximately $0.12 per summary for DESCRIBE and $0.05 for VG-prompt. We note that this cost covers a complete multi-sentence assessment (averaging five sentences) with structured rationales, whereas prior work typically evaluates shorter summaries (Si et al., 2026). The decomposition step is performed once per summary and can be reused across different inference configurations or backbone models (as demonstrated by the DESCRIBE$^{\text{gpt5}}$ variant), amortizing this cost. Lastly, the RST parser introduces minimal additional overhead and can process the full VERIGRAY dataset by predicting parsed RST trees within two minutes on a single GPU once trained.

[19] https://lambdalabs.com/service/gpu-cloud

## 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 8: FineSURE and VG-prompt prompts for faithfulness evaluation.

We also noted that in the original implementation of VG-prompt (Ding et al., 2025), the authors conducted independent sentence-level evaluation, which in turn led to 2,044 inferences with no significant performance gain compared to our re-implementation.

## F Significance Testing

Recent work leveraging LLMs as baselines often ignores the importance of significance testing (Seo et al., 2025; Si et al., 2026) or provides vague descriptions of the significance-testing setup (Liu et al., 2026). Following Koehn (2004), we conduct a paired bootstrapping test between our DESCRIBE and the baseline VERIGRAY. In detail, we use paired bootstrap resampling over documents;[20] and average performance across multiple runs within each resample to account for model stochasticity. In each bootstrap sample, we compute the metric difference between two systems on the resampled set; this is repeated $5,000$ times, and significance is determined from the bootstrap distribution at $p < 0.05$. We did not run tests on the $L_{rank}$, since the computation operates on large-scale sentence-pair comparison ($N^2$ complexity and average of more than one second per run, which is time-consuming.)

## G Additional Analysis

### G.1 Decomposition Error Analysis

Quantitatively, we analyzed the number of subclaims generated by each method. On average, there are 13.3, 12.0, and 19.3 decomposed claims generated by our approach, Fables, and VeriFact, respectively. We also observed a significantly higher claim-per-sentence ratio for VeriFact, which achieved an average of 5.1, compared to 2.7 for our approach and 2.6 for Fables. This denotes that the last approach might suffer from the over-decomposition issue. On the qualitative perspective, as shown in Figure 7, our approach obtained a lower rate of contextual information omission and meaning alternation errors (in total 13.3%) compared to Fables with 20% errors. A closer analysis of the assessment results shows that our approach effectively preserves the key logical relationship across claims (i.e, the causal relation in example 1 of Table 5) through the discourse modeling.

[20]This is to preserve the connections within the same summary.

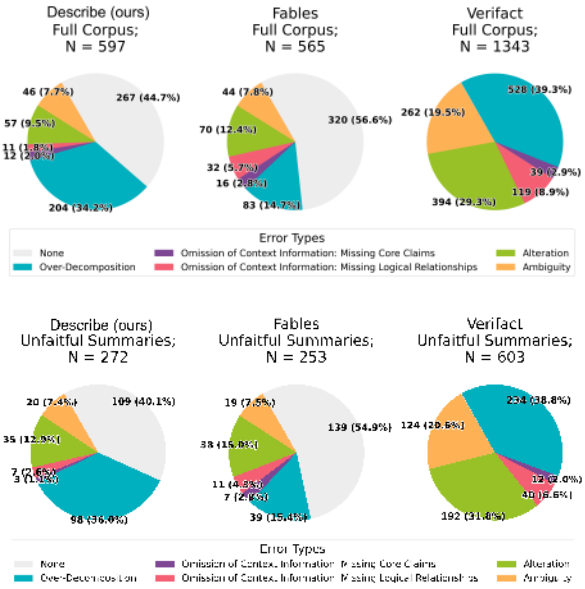



Figure 7: Decomposition error distribution across approaches (including DESCRIBE): full corpus (top) vs. summaries with $\geq$ 1 human-labeled error (bottom).

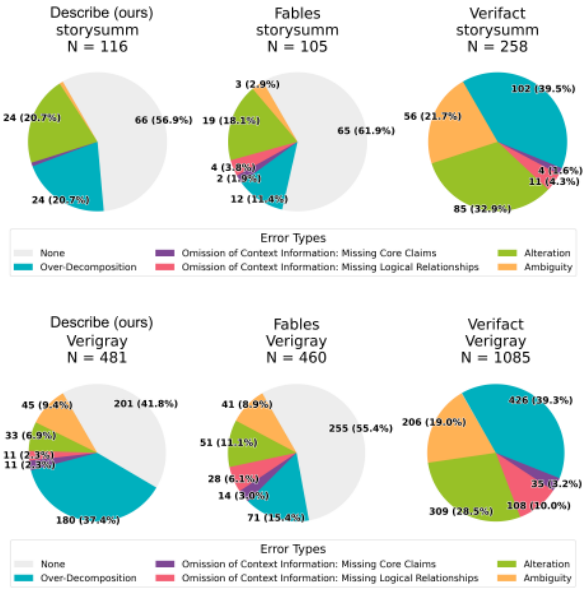



Figure 8: Decomposition error distribution between our approach and prior approaches for each benchmark dataset.

### G.2 Ablation Setup

To examine the effects of different claims in Section 6.2, we modify the final portion of Figure 18 by replacing *### Summary Claims and Relation Table: CLAIM_PLACEHOLDER* with *### Summary Claims without relation table: NEW_CLAIM_PLACEHOLDER*. For the bottom

| ID | Claim | BAcc | MCC | Hallu. R.↑ | Hallu. P↑ | Hallu. F1↑ | $L_{rank}$ ↓ |
|---|---|---|---|---|---|---|---|
| *Fixed structured inference, varying decomposition* | | | | | | | |
| 1 | DESCRIBE$^{gpt5}$ | 76.8 | 0.65 | 55.0 | 87.8 | 67.7 | 35.8 |
| 2 | w/o rel. | 76.1 | 0.64 | 54.2 | 84.6 | 65.6 | 36.0 |
| 3 | Fables | 75.8 | 0.63 | 53.3 | 85.8 | 65.7 | 36.1 |
| 4 | Verifact | 75.5 | 0.64 | 52.9 | 85.0 | 65.2 | 36.2 |
| 5 | Original Sent | 75.8 | 0.63 | 53.7 | 84.1 | 65.5 | 36.1 |
| *Fixed VG-prompt baseline inference* | | | | | | | |
| 6 | DESCRIBE | 75.8 | 0.64 | 52.8 | 86.0 | 65.8 | 35.0 |
| 7 | Origin sent | 73.9 | 0.60 | 49.5 | 85.4 | 62.7 | 35.6 |

Table 9: Ablation study analyzing the effects of decomposition and inference strategies on VERIGRAY, using Gemini-3-Flash backbone.

part with VG-prompt baseline prompting, we modify the prompt to add the instruction: *"please also consider the decomposed claims when doing your evaluation: {row['gpt5-pred']}. You need to leverage the relation fields that are used to connect different subclaims. For instance, if a sentence is split into two claims with a causal relation, you need to check both claims' correctness and verify that the relation is preserved and aligned with the original document.*

Table 9 reports the ablation results using Gemini-3-Flash as the backbone, with a similar trend as reported in Section 6.2.

### G.3 Model Gains on Stratified Dataset

We hypothesize that our discourse-inspired approach may have different levels of benefits over summary sentences with varying levels of complexity. One way to evaluate sentences is through studying the existence of explicit discourse connectives (Ko and Li, 2020). We thus split the datasets into two groups based on existences explicit discourse connectives detected by a predictor model. In the following section, we provide the implementation details on the discourse connective detectors and further discuss our main findings.

**Connective Detector.** The task of discourse connective detection (identifying tokens in the sentence as starting, being inside, or outside a discourse connective) has become a standard task in the Discourse Relation Parsing and Treebanking (DISRPT) shared task (Braud et al., 2023, 2025). We retrained one of the participating models from DISRPT 2023, HITS (Liu et al., 2023b),[21] on the `eng.pdtb.gum` dataset;[22] as no pretrained checkpoint was publicly available. The model consists of a BERT encoder followed by a BiLSTM and a conditional random field (CRF) layer to predict token labels under the BIO tagging scheme. Our retrained model achieved an F1 score of 87.59 on the test set for the discourse connective detection task, matching the top result in the official shared task report (Braud et al., 2025). Although the model is not perfect, we find that it can reliably extract discourse connectives from summary texts for downstream analysis.

**Connectives Extraction and Data Split.** Leveraging the detector, we process summaries of VERIGRAY to extract discourse connectives and further split the dataset (sentence-level) into two groups: those containing no explicit discourse connective (n=1015), and those containing discourse connectives such as *"and"*, *"because"*, *"but"* (n=1029).

**DESCRIBE's Gains over Baselines.** Figure 9 shows the complete improvements comparing DESCRIBE to the baselines, with three reported metrics on two sets of backbones.

### G.4 Removing RST Tree for DESCRIBE

We experiment with a variant of DESCRIBE in which, during the decomposition phase, the model is instructed to perform discourse parsing directly from demonstrative examples rather than being provided with the discourse tree predicted by a fine-tuned parser. As shown in Table 10, without a parser-inferred discourse tree as input, the downstream performance results in a performance drop, although the variant still consistently outperforms the baseline approach.

Similarly, when analyzing performance across stratified data splits, as in Appendix G.3, we observe that the primary performance drop from replacing the parser-generated RST trees occurs in subsets of sentences with diverse discourse connectives (Table 11).

## H Rationale Evaluation Protocol

The study of faithful model explanation has raised more attention with the emergence of large language models (Lyu et al., 2024). We begin by reviewing the literature on existing work that proposes different evaluation dimensions of rationales.

[21] https://github.com/liuwei1206/disrpt2023

[22] https://github.com/disrpt/sharedtask2025/tree/master/data

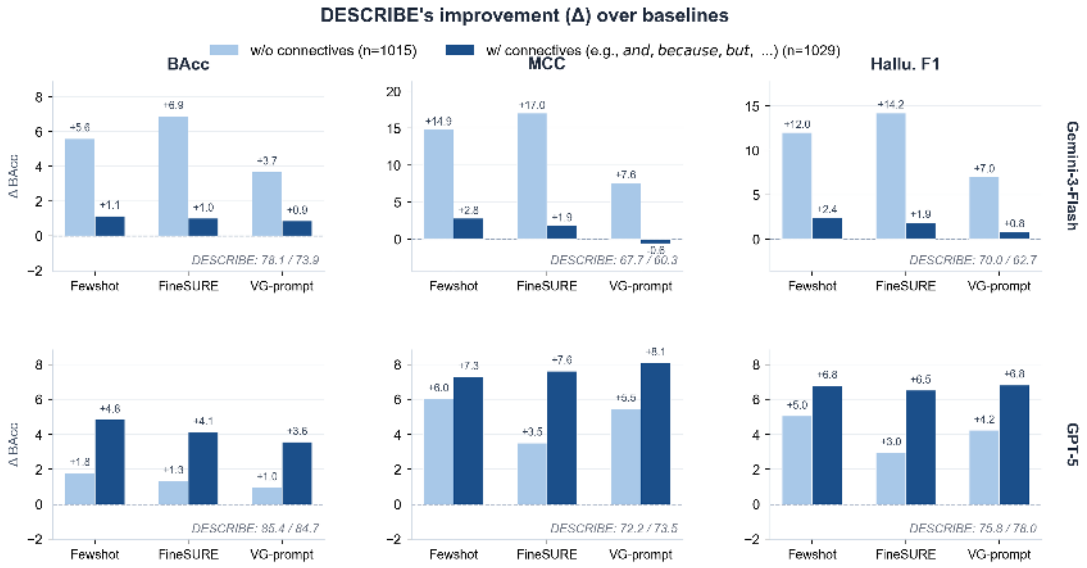



Figure 9: **DESCRIBE's advantage over baselines on stratified sentence groups in VERIGRAY.** Sentences are split based on the presence of predicted discourse connectives using a trained discourse parser. Bars show Δ show absolute differences. DESCRIBE's absolute scores are shown in *italics*.

| Method | BAcc | MCC | Hallu. R.↑ | Hallu. P↑ | Hallu. F1↑ | $L_{rank}$ ↓ |
|---|---|---|---|---|---|---|
| DESCRIBE w/ parse-generated tree | 84.9 | 0.73 | 73.2 | 81.5 | 77.1 | 30.8 |
| w/o parsed tree | 83.2 | 0.68 | 70.7 | 76.3 | 73.4 | 31.5 |
| Baseline (VG-prompt) | 82.5 | 0.64 | 71.0 | 69.8 | 70.4 | 31.6 |

Table 10: Comparison of the baseline approach, DESCRIBE, with a variant of DESCRIBE that does not leverage the discourse parser-generated RST tree on VERIGRAY benchmark with GPT-5.

Seo et al. (2025) introduces three binary assessment criteria: *Completeness*, *logical Coherence*, and *Faithfulness* to assess the coverage, logical flow, and validity of the rationale contexts. Similarly, Si et al. (2026) propose *Informativeness*, *Readability*, and *Helpfulness* to evaluate the similar dimension. LITEX (Hong et al., 2025) introduces an eight-way fine-grained taxonomy, which can be divided into text-based reasoning and world knowledge-based reasoning to assess rationales on NLI pair assessment. However, none of the aforementioned work utilizes the human rationales provided in the benchmarking corpora. Another overlooked problem, as found in our analysis in Section 3, is the lack of discourse modeling checking. To this end, we additionally propose two more dimensions to assess the human rationale alignment and the capturing of discourse relations. Inspired by prior work (Miao et al., 2024), which applied a question–answer–based approach to evaluate the faithfulness of discourse understanding in different models, we design a prompt that explicitly presents discourse relations with examples and further validate whether such relations are correctly identified in the generated rationales.

We build our evaluation prompt on top of the template of Seo et al. (2025), and include all prompting details in Section H.1.

For Human Rationale Alignment and Discourse Awareness, we retain only instances with valid human-annotated rationales. This results in 80+ unfaithful sentences from STORYSUMM and 564 sentences from VERIGRAY, where either the "note" or "rationale" field is non-empty, and the annotation label falls into one of the following categories: generally-supported", fabricated", or "contradicting".

### H.1 Evaluation Prompts

We include the full evaluation prompts for **Completeness**, **Reasoning Quality**, **Human Rationale Alignment**, and **Discourse Awareness** in Figures 12, 13, 14, and 15 respectively.

| Connective Split | System | BAcc. | MCC | Hallucination F1 |
|---|---|---|---|---|
| w/o connective | DESCRIBE w/ parsed tree | **85.44** | **72.23** | **75.83** |
| | DESCRIBE w/o parsed tree | 84.38 | 68.41 | 72.65 |
| | Δ | +1.06 | +3.82 | +3.18 |
| w/ connectives (and, but, because, when ...) | DESCRIBE w/ parsed tree | **84.66** | **73.52** | **78.03** |
| | DESCRIBE w/o parsed tree | 82.36 | 68.34 | 73.86 |
| | Δ | +2.36 | +5.18 | +4.14 |

Table 11: **Performance comparison over splits of VERIGRAY between DESCRIBE using parsed RST trees and not**. Values are percentages (%). Δ indicates absolute improvement.

| System | $n_1$ | Compl. | Reas. | $n_2$ | H.Align | Disc. (↑ / ↓ / –) |
|---|---|---|---|---|---|---|
| *STORYSUMM* | | | | | | |
| DESCRIBE | 497 | 89.7 | 97.0 | 42 | 92.9 | 85.7/11.9/2.4 |
| VG-prompt | 488 | 73.8 | 96.1 | 53 | 75.5 | 71.7/26.4/1.9 |
| FineSURE | 476 | 55.9 | 90.8 | 58 | 72.4 | 60.3/37.9/1.7 |
| FaithLens | 481 | 84.4 | 84.8 | 32 | 21.9 | 65.6/31.2/3.1 |
| *VeriGray* | | | | | | |
| DESCRIBE | 1741 | 89.7 | 98.4 | 343 | 91.0 | 89.2/7.6/3.2 |
| VG-prompt | 1746 | 84.0 | 94.0 | 357 | 85.4 | 70.6/21.0/8.4 |
| FineSURE | 1734 | 65.2 | 97.6 | 330 | 81.8 | 61.5/30.3/8.2 |
| FaithLens | 1581 | 91.9 | 90.4 | 193 | 68.9 | 48.7/36.3/15.0 |

Table 12: **Rationale evaluation (%) on correctly predicted instances (GPT-5-mini as evaluator).** $n_1$: instances where $\hat{y} = y_{\text{gold}}$; $n_2$: subset of $n_1$ where human rationales are available. Disc. reports proportions of correct (↑) / incorrect (↓) / not applicable (–) discourse identification.

| System | $n_1$ | Compl. | Reas. | $n_2$ | H.Align | Disc. |
|---|---|---|---|---|---|---|
| ***GPT-5-mini*** | | | | | | |
| *STORYSUMM* | | | | | | |
| DESCRIBE | 83 | 88.0 | 96.4 | 62 | 46.8 | 93.5 |
| VG-prompt | 92 | 65.2 | 91.3 | 51 | 45.1 | 74.5 |
| FineSURE | 104 | 39.4 | 78.8 | 46 | 32.6 | 65.2 |
| FaithLens | 99 | 72.7 | 72.7 | 72 | 2.8 | 59.7 |
| *VeriGray* | | | | | | |
| DESCRIBE | 263 | 88.6 | 96.6 | 221 | 41.6 | 70.6 |
| VG-prompt | 282 | 79.1 | 84.8 | 207 | 24.6 | 51.7 |
| FineSURE | 304 | 63.2 | 94.4 | 231 | 24.7 | 42.9 |
| FaithLens | 463 | 79.5 | 70.2 | 371 | 16.7 | 38.5 |
| ***Gemini-3-Flash*** | | | | | | |
| *STORYSUMM* | | | | | | |
| DESCRIBE | 83 | 92.8 | 94.0 | 62 | 48.4 | 96.8 |
| VG-prompt | 92 | 75.0 | 91.3 | 51 | 43.1 | 88.2 |
| FineSURE | 104 | 46.2 | 77.9 | 46 | 34.8 | 82.6 |
| FaithLens | 99 | 81.8 | 58.6 | 72 | 4.2 | 73.6 |
| *VeriGray* | | | | | | |
| DESCRIBE | 263 | 92.4 | 98.1 | 221 | 39.8 | 83.7 |
| VG-prompt | 282 | 88.7 | 91.8 | 207 | 30.9 | 64.7 |
| FineSURE | 304 | 66.4 | 93.4 | 231 | 16.9 | 58.9 |
| FaithLens | 463 | 86.6 | 65.0 | 371 | 15.1 | 50.4 |

Table 13: **Rationale evaluation (%) on incorrectly predicted instances.** $n_1$: instances where $\hat{y} \neq y_{\text{gold}}$; $n_2$: subset of $n_1$ where human rationales are available.

### H.2 Additional Automatic Assessment with GPT-5-mini

Table 12 presents the rationale evaluation results across four dimensions on correctly predicted instances, using GPT-5-mini (`gpt-5-mini-2025-08-07` with reasoning effort set as *high*) as the backbone model for evaluators. We observe similar patterns as reported in Section 6.3.

### H.3 Analysis on Disagreed Prediction Instances

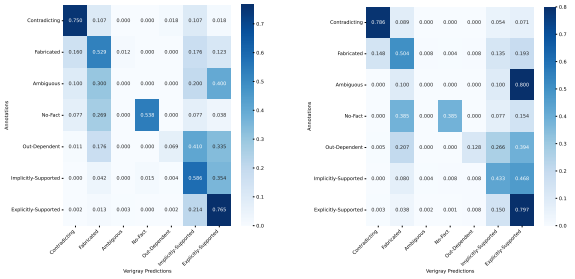
Table 13 shows the evaluation results on instances where the system's predicted label does not match the human labels. DESCRIBE consistently obtains the highest discourse awareness score, and preserves a noteworthy H.Align score (36.6% on VERIGRAY using GPT-5-mini) compared to other baselines such as FineSURE and FaithLens. This suggests that when the final prediction is incorrect, DESCRIBE's reasoning process remains more faithful to human judgment and discourse structure, potentially highlighting a gap between label prediction and the rationale.

### H.4 Manual Evaluation

To ensure diverse coverage of prediction-rationale alignment patterns, we categorize all instances into four groups based on cross-system prediction correctness and automatic rationale scores: (1) *all good* – all systems predict correctly with high rationale scores, (2) *label-rationale mismatch* – correct prediction but at least one rationale dimension fails, or incorrect prediction despite sound reasoning, (3) *system diverge* – systems differ in both prediction and rationale quality, and (4) *all poor* – all systems produce low-quality rationales. We oversample categories (2) and (3) as they are most diagnostic for evaluating rationale quality, while including a smaller proportion of the category (1). This yields 50 instances, each evaluated independently by two annotators across all three GPT-5-based model variants (anonymized and randomly ordered), producing 300 total judgments (four dimensions per judgment).

Before participants begin to make judgments, we describe the principles of our design in detail

and ensure that each participant correctly understands the principles. The first author reviewed 10 held-out examples with each annotator individually and carried out a one-hour training session, ensuring that annotators were comfortable with the instructions. We provide access to the evaluation dimension and details in a separate doc file. Two annotators manually labeled the data independently, resolving disagreements with adjudication. They also consent that their data may be used for quantitative analysis and included in the paper. The payment also ensures that workers were paid ($10-$13 per hour) beyond the minimum wage. Both participants hold master's degrees in computer science/information science and have been exposed to natural language processing tasks before. The screenshot of the annotation interface is included in Figure 11.



(a) DESCRIBE (b) VG-prompt baseline.

Figure 10: Confusion matrices for two model predictions of fine-grained labels on VERIGRAY.

The inter-annotator agreement (IAA), measured using Cohen's Kappa, is 0.91 for H.Align, 0.80 for Disc., and 0.41 for Completeness. Although the Completeness score falls within the moderate agreement range, manual inspection indicates that this is primarily driven by systematic differences in annotation strategy: one annotator adopted a pairwise comparison approach when evaluating three system outputs, consistently preferring the anonymized DESCRIBE version and assigning scores of 0s to the others.

# I Error Analysis Supplementary Materials

## I.1 Fine-grained classification

The fine-grained macro-F1s on VERIGRAY for DESCRIBE and the VG-prompt analyzed in Section 6.1 are 44.2 and 42.6, respectively. Figure 10 visualizes the confusion matrices for the two approaches.

## I.2 Additional Errors

Besides the three most frequent error types, we document additional errors and provide examples in Table 15: (1) *Overlooked Annotation Errors*, arising from the label aggregation protocol in STORYSUMM; (2) *Subjective Wording Errors*, where summary wording overlaps with the source but allows alternative interpretations; and (3) *Other Errors*, covering infrequent and nuanced cases that do not fit the primary categories.

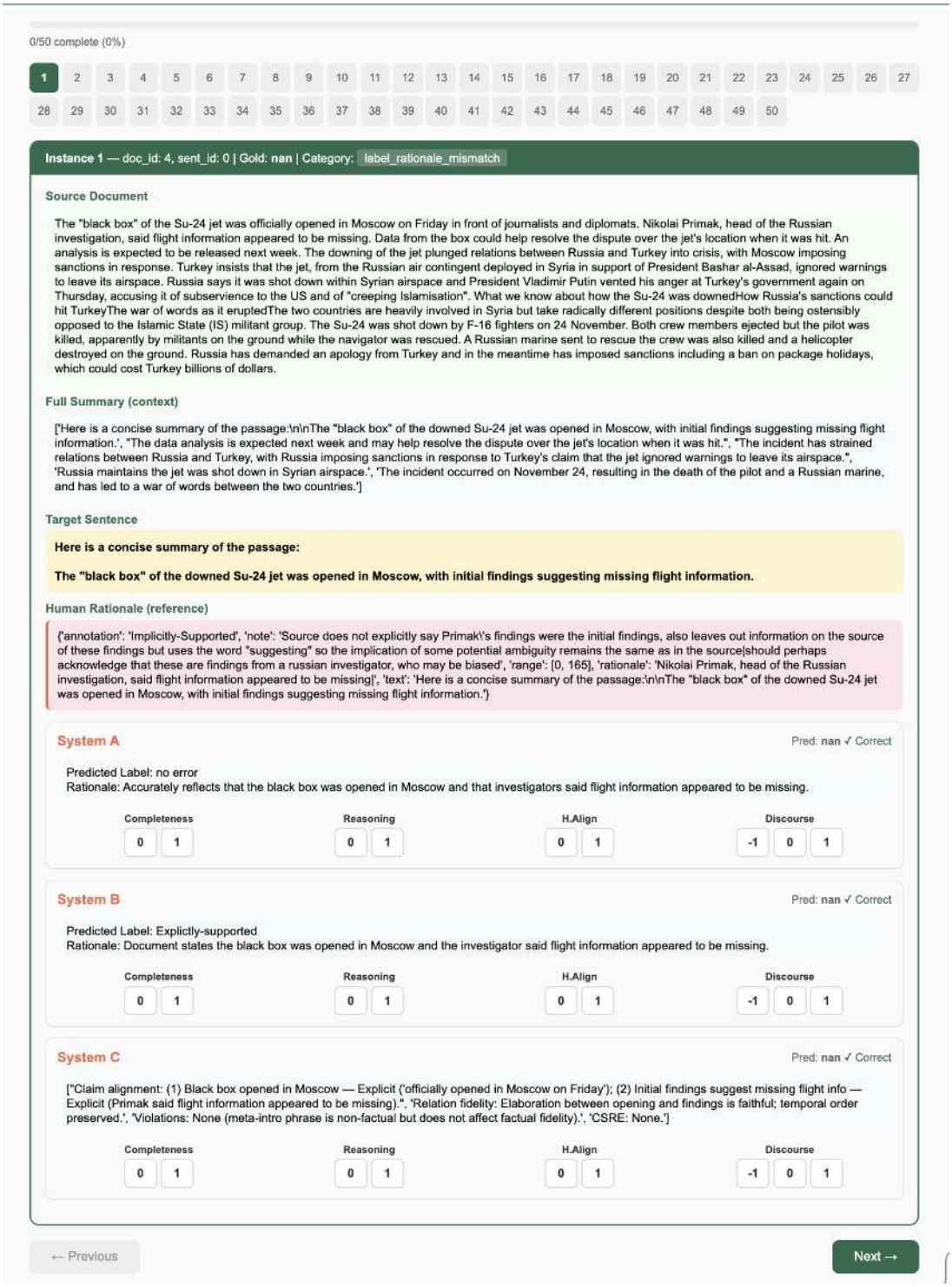



Figure 11: Human Assessment Interface.

| Source & Summary & Human Annotation | Our System Prediction | Manual Error Category & Automatic Rationale Assessment (§6.3) |
|---|---|---|
| "Obviously … the more games we can win the more confidence we can get from that we can hopefully go on a run of games and push up the table and try to get the club back to where it has been," added Clarkson.<br><br>**Summary Sentence:** St Mirren, currently eighth in the Championship, aim to climb the table with Clarkson's help.<br>**Gold Faithfulness Label:** *Fabricated*<br>**Human Rationale:** Note: this is Clarkson's aim. | {**Predicted Label:** Generally-Supported<br>**"Rationale":** [ Explicit: "St Mirren are eighth in the Championship" is directly stated in the source.", "Aim: Clarkson states that they hope to "push up the table," which supports the claim that the team aims to improve its standing.", "with Clarkson's help" introduces a causal linkage that is not explicitly stated; it is weakly entailed since he is both a team member and the speaker, but this contribution is not directly asserted. This constitutes a minor causal overreach.", CSRE note: The use of his past achievements as evidence for the team's objective is not stated in the source. Overall, the claim is generally supported." ]} | **Matched Rationale but Wrong label**<br>**Assessment:** Rationale Align 0<br>***Feedback***: The model identifies the same key evidence as the human rationale ("aim" to climb the table is voiced by Clarkson himself). However, human rationale uses the attribution of the aim to Clarkson (rather than the club) as the basis for a "fabricated" label, the model's reasoning concludes that the sentence is "overall generally supported". |
| Centre-back Anderson, 27, has cancelled his Barrow contract by mutual consent to join their National League rivals. The ex-Aldershot and Exeter player, who has also played in Italy, could feature against his old club on 18 February.<br><br>**Summary Sentence:** The ex-Aldershot and Exeter player, with prior experience in Italy, may debut against his former club on 18 February.<br>**Gold Faithfulness Label:** *Fabricated*<br>**Human Rationale:** Note: "debut" is not mentioned. | { **Predicted Label:** Generally-Supported,<br>**"Rationale":** ["C4 (ex-Aldershot and Exeter) – Explicitly supported.",<br>"C5 (has played in Italy) – Explicitly supported.",<br>"C6 (may debut vs. could feature on 18 February vs. former club) – Entailed: the source states "could feature against his old club on 18 February"; the summary's phrase "may debut" introduces a first-appearance implication that is not explicitly stated but is reasonably inferred given the context of a new signing.",<br>"Relation fidelity: Elaboration and attribute relations are preserved; the contingency on joining is implicitly supported by the source context; temporal alignment remains consistent.",<br>"Violations: Minor factual nuance – the term 'debut' introduces a first-appearance implication that is not explicitly stated in the source." ]} | **Over-interpretation in Ratioanle**<br>**Assessment:** Rationale Align 0<br>***Feedback:*** The core issue is the addition of a first-appearance nuance "debut". The model, however, treats "may debut" as reasonably implied by "could feature" and marks it as entailed, explicitly contradicting the human rationale. |
| Nationwide over the course of the decade, the rate of binge drinking among women increased more than seven times the rate among men.<br><br>**Summary Sentence:** Binge drinking rates increased more than seven times faster among women than men nationwide.<br>**Gold Faithfulness Label:** *Fabricated*<br>**Human Rationale:** Note: over the course of a decade. | {**Predicted Label:** Explicitly-Supported<br>**"Rationale":** [ Explicit support: "Nationwide over the course of the decade, the rate of binge drinking among women increased more than seven times the rate among men.", Relation fidelity: This statement functions as evidence for the women-driven increase described in S1 and aligns with the Explanation/Evidence relation in the source.<br>"Violations: Minor wording variation ("times faster" vs. "times the rate"), but the content is equivalent.", "CSRE: The alignment between S1 and S2 is preserved; no relational mismatch is observed." ] } | **Missing Fine-grained Details**<br>**Assessment:** Discourse 0<br>***Feedback:*** The target sentence embeds a comparison (rate of increase for women vs men) coupled with a temporal framing ("increased ... nationwide over the course of the decade"). The model's rationale correctly cites the texts... However, it fails to explicitly identify the central comparison relation (and only partially acknowledges temporality) |

Table 14: **Examples of DESCRIBE's rationale generation with automatic assessment and human-categorized error types.** Each row illustrates a distinct failure mode: (1) *Matched Rationale but Wrong Label* – the model selects the same evidence as human annotators but draws a different conclusion due to different attribution interpretation; (2) *Over-interpretation* – the model treats implied information as entailed, contrary to the human judgment; (3) *Missing Fine-grained Details* – the model verifies surface content but overlooks key discourse relations (e.g., the temporal framing) and the comparison structure. CSRE denotes cross-sentential relation evaluation.

| Source & Summary & Human Annotation | Our System Prediction |
|---|---|
| It started right on Kristen's birthday (21st) ... While Kristen's family was very well off now, for the first 12 years of her life they lived in extreme poverty.<br><br>**Summary Sentence**: Kristen realizes that her family's improved fortune over the last 12 years likely came as the result of a sinister deal made by her mother, who vanished around that time.<br>**Human Faithfulness Label**: 1 (faithful) | **Predicted Label:** *Contradicting*<br>**Model Rationale:** [ "Partially supported inference: Father says the mother "sold us out," and the family's luck turned almost overnight; the mother vanished shortly after" this supports a likely bargain by the mother (hedged)",<br>"Contradiction in details: The text states the first 12 years were in extreme poverty; the improved fortune began after age 12, meaning 9 years of prosperity up to age 21, not "over the last 12 years.",<br>R̈elation fidelity: C6 (temporal) is supported; C4-5 (background) is undermined by the incorrect duration.̈], |
| I've always had good hearing but that day, not only was I hearing very well, I was agitated by what I was hearing, no matter what it was. If it was even louder than normal speaking voice, I would get super flustered.<br><br>**Summary Sentence:** One day, the narrator experiences heightened hearing and becomes agitated by loud noises.<br>**Human Faithfulness Label**: 0 (unfaithful) | **Predicted Label:** Explicitly-supported<br>**Model Rationale:** :[ Source states: "that day … I was hearing very well" (heightened hearing); Source states agitation at sounds: "I was agitated by what I was hearing … If it was even louder than normal speaking voice, I would get super flustered."; Temporal framing ("one day/that day") matches; No added causality or attribution shift; JOINT listing of co-occurring states matches. ] |

Table 15: **Additional qualitative examples of DESCRIBE's rationale generation with human-categorized error types.** Each row illustrates a distinct failure mode: (1) *Annotation Error* and (2) *Subjective Wording Error*.

## Completeness

You are a fair judge assistant tasked with providing clear, objective feedback based on specific criteria, ensuring each assessment reflects the absolute standards set for performance. Another model (fact verifier) was tasked with assessing whether a source document supports a given summary sentence. The model then generated an explanation/rationale for its conclusion. Your role is to evaluate the quality of the rationale along specified task dimensions.

**### Task Description**
Whether the fact verifier explicitly checks all critical components of the given statement, regardless of verification accuracy. The output format should be as follows: Feedback: (provide a detailed evaluation based on completeness) [RESULT] (an integer, either 0 or 1).

Do not include any additional opening statements, explanations, or conclusions beyond the specified format.

**### Evaluation Policy**
The evaluation of the rationale must focus on the verification of the target summary sentence. The full summary is provided only for contextual reference. Before assigning a score, evaluate the response by checking:
1. Thoroughness of Verification
2. Completeness of Checking
3. Avoidance of Partial Verification

If any of these checks fail, the response should be rated 0.

**Document:**
DOCUMENT_PLACEHOLDER

**Full Summary (for context only):**
SUMMARY_PLACEHOLDER

**Target Summary Sentence:**
STATEMENT_PLACEHOLDER

**Model Response:**
MODEL_RESPONSE_PLACEHOLDER

**### Score Rubric (Completeness Evaluation)**
**Score 0 (Low Completeness):** The response does not verify all critical components of the statement. It either skips key parts, provides an incomplete assessment, or focuses on only a subset of the facts presented in the statement. The response is not a thorough verification and cannot be considered fully precise.
**Score 1 (High Completeness):** The response explicitly checks all critical aspects of the statement. Each part of the statement is verified without omission, and no essential details are overlooked. The response ensures that the verification is complete, without introducing unrelated details.

Please review the model response above carefully. Then, output your evaluation in the following format:
**Feedback:** <*detailed feedback explaining your evaluation*> **[RESULT] <0 or 1>**

Figure 12: Prompt for assessing completeness of rationale.

## Reasoning Quality

You are a fair judge assistant tasked with providing clear, objective feedback based on specific criteria, ensuring each assessment reflects the absolute standards set for performance. Another model (fact verifier) was tasked with assessing whether a source document supports a given summary sentence. The model then generated an explanation/rationale for its conclusion. Your role is to evaluate the quality of the rationale along specified task dimensions.

**### Task Description**
Whether each inference step in the fact verifier's reasoning is logically sound and justified; specifically, evaluate the internal consistency and validity of rationales.

The output format should be as follows:
**Feedback:** (provide a detailed evaluation based on reasoning quality) **[RESULT]** (an integer, either 0 or 1)

Do not include any additional opening statements, explanations, or conclusions beyond the specified format.

**### Evaluation Policy**
The evaluation of the rationale must focus on the verification of the target summary sentence. The full summary is provided only for contextual reference. Before assigning a score, evaluate the response by checking:
1. Accuracy of Logical Steps
2. Avoidance of Incorrect Inferences
3. Consistency in Fact Interpretation
4. Sound Justification for Verification

If any of these checks fail, the response should be rated 0.

**Document:**
DOCUMENT_PLACEHOLDER

**Full Summary (for context only):**
SUMMARY_PLACEHOLDER

**Target Summary Sentence:**
STATEMENT_PLACEHOLDER

**Model Response:**
MODEL_RESPONSE_PLACEHOLDER

**### Score Rubric (Reasoning Quality Evaluation)**
**Score 0 (Low Reasoning Quality):** The model's reasoning is flawed, misinterprets the document, or applies incorrect logic when verifying the statement. The response contains unsupported inferences, misreads facts, distorts meaning, or jumps to a conclusion that is not justified by the evidence. The model may misapply classification rules, make unwarranted logical assumptions, or draw causality that is not present in the document. If the rationale is internally consistent but based on a misreading of the source, the score should be 0. If the response demonstrates any fundamental flaws in reasoning, the score must be 0.

**Score 1 (High Reasoning Quality):** The model accurately applies logical reasoning and correctly interprets the document when verifying the statement. The reasoning properly connects facts, avoids misinterpretations, and maintains logical validity from extraction to conclusion. The model does not introduce unjustified assumptions or incorrectly infer relationships between facts, and it provides a structured, step-by-step justification that correctly supports its classification.

Please review the model response above carefully. Then, output your evaluation in the following format:
**Feedback:** <*detailed feedback explaining your evaluation*> **[RESULT] <0 or 1>**

Figure 13: Prompt for assessing reasoning quality of a rationale.

## Human Rationale Alignment

You are a fair judge assistant tasked with providing clear, objective feedback based on specific criteria, ensuring each assessment reflects the absolute standards set for performance. Another model (fact verifier) was tasked with assessing whether a source document supports a given summary sentence. The model then generated an explanation/rationale for its conclusion. Your role is to evaluate the quality of the rationale along specified task dimensions.

**### Task Description**
Whether the model's reasoning captures the essential justification(s) provided in the human-annotated rationale.

The output format should be as follows:
**Feedback:** (provide a detailed evaluation based on human rationale alignment) **[RESULT]** (an integer, either 0 or 1)

Do not include any additional opening statements, explanations, or conclusions beyond the specified format.

**### Evaluation Policy**
The evaluation of the rationale must focus on the verification of the target summary sentence.
Before assigning a score, evaluate the response by checking:
1. Whether the model identifies the same key evidence or facts highlighted in the human rationale
2. Whether the model's reasoning reflects the same core justification logic
3. Whether the model avoids contradicting the human rationale

If any of these checks fail, the response should be rated 0.

**Document:**
DOCUMENT_PLACEHOLDER

**Full Summary (for context only):**
SUMMARY_PLACEHOLDER

**Target Summary Sentence:**
STATEMENT_PLACEHOLDER

**Human Label and Rationale:**
HUMAN_LABEL_RATIONALE

**Model Response:**
MODEL_RESPONSE_PLACEHOLDER

**### Score Rubric (Human Rationale Alignment)**
**Score 0:** The model's reasoning does not reflect the core justification provided in the human rationale. It may rely on unrelated evidence, omit key reasoning, or contradict the human explanation.
**Score 1:** The model's reasoning captures the essential justification expressed in the human rationale, relying on the same key evidence or logical basis.

Please review the model response above carefully. Then, output your evaluation in the following format:
**Feedback:** <*detailed feedback explaining your evaluation*> **[RESULT] <0 or 1>**

Figure 14: Prompt for assessing the alignment of human rationale.

## Discourse Relation Awareness

<SAME ROLE INSTRUCTION AS PRIOR TASKS>

**### Task Description**

Whether the fact verifier's rationale correctly recognizes and evaluates discourse relations present in the target summary sentence when such relations affect its truth conditions. Discourse relations include (but are not limited to): causal, contrastive, concessive, conditional, temporal, purpose, comparison, attribution, or scope-related relations.

**### Output Format**

**Feedback:** (provide a detailed evaluation based on discourse relation awareness) **[RESULT]** (an integer, either -1, 0, or 1)
Do not include any additional opening statements, explanations, or conclusions beyond the specified format.

**### Evaluation Policy**

The evaluation of the rationale must focus on the verification of the target summary sentence. The full summary is provided only for contextual reference.

**Step 1: Determine Applicability**

First determine whether the target summary sentence contains explicit or implicit discourse relations that are necessary for its meaning or truth conditions. A discourse relation is considered present if the sentence includes the following contents explicitly or implicitly:

1. Causal structure (e.g., because, resulting in, leading to)
2. Contrast or concession (e.g., although, however, despite)
3. Conditional structure (e.g., if, unless)
4. Temporal ordering (e.g., before, after, during)
5. Purpose (e.g., in order to, to + verb)
6. Attribution (e.g., said, claimed, according to)
7. Comparison or scope constraints (e.g., more than, only, at least)

If no such discourse relation is present, assign **Score -1** and briefly state that no discourse relation is involved.

**Step 2: If Discourse Relation Is Present**

Before assigning a score, evaluate whether:

1. The rationale explicitly recognizes the discourse relation.
2. The rationale preserves the correct relation type (e.g., cause vs. contrast).
3. The rationale maintains correct directionality or scope (e.g., cause $\rightarrow$ effect, before $\rightarrow$ after, attribution vs. assertion).
4. The rationale evaluates document evidence in a way that directly addresses the discourse relation, not merely entity overlap.

If any of these checks fail, the response should be rated **0**. Otherwise, score **1**.

**### Score Rubric (Discourse Relation Awareness Evaluation)**

**Score -1 (No Discourse Relation embedded):** The target summary sentence does not contain explicit or implicit discourse relations affecting its truth conditions. This criterion is not applicable.
**Score 0 (Incomplete/Incorrect Discourse Awareness):** The rationale ignores a discourse relation present in the sentence, misclassifies the relation type, reverses its directionality, mishandles attribution or scope, or evaluates only surface-level entity overlap without assessing the relational structure necessary for truth conditions.
**Score 1 (Correct and Complete Discourse Awareness):** The rationale correctly identifies and evaluates any discourse relation present in the sentence, preserves its type and directionality, and grounds the verification in document evidence that directly addresses the relational structure.

**Document:**
DOCUMENT_PLACEHOLDER

**Full Summary (for context only):**
SUMMARY_PLACEHOLDER

**Target Summary Sentence:**
STATEMENT_PLACEHOLDER

**Human Rationale:**
HUMAN_RATIONALE_PLACEHOLDER

**Model Response:**
MODEL_RESPONSE_PLACEHOLDER

Please review the model response above carefully. Then, output your evaluation in the following format:
**Feedback:** <*detailed feedback explaining your evaluation*> **[RESULT] <-1, 0, or 1>**

Figure 15: Prompt for assessing awareness of discourse relations.

## RST Atomic Claim Decomposition Instructions

You are an expert in Computational Linguistics, specializing in Rhetorical Structure Theory (RST). Your goal is to deconstruct a summary text into "Atomic Claims.", each of which will then be passed to a human or an LLM evaluator who will read the original document that the summary is based on and verify if the claim is true or not, and further build the intra and inter relation graphs across claims to represent the document better. Each atomic claim must be fully understandable **without any external context** (e.g., all entities must be referred to by name instead of pronouns), and claims should be situated within relevant **temporal, locational, and causal context** wherever possible.

**## RST Relation Definitions**
In the following definitions:
- **W** = Writer (or speaker)
- **R** = Reader (or hearer)
- **S** = Satellite
- **N** = Nucleus

RST relations are categorized as either mononuclear (comprising a nucleus and a satellite span) or multinuclear (comprising two or more nuclei spans).
| Relation | Description |
| **Adversative** | Connects discourse units where some incompatibility is highlighted. |
| **Attribution** | S provides information about the source of information in N. |
| **Causal** | S and N form a cause–result pair. |
| **Context** | S provides circumstances (often spatio-temporal) under which N applies. |
| **Contingency** | N occurs depending on S. |
| **Elaboration** | S gives additional information about N. |
| **Evaluation** | S provides an assessment of N by W. |
| **Joint** | Connects multiple, non-contrasting units of equal prominence. |
| **Explanation** | S supports N by explaining its utterance. |
| **Mode** | S explains how N happens. |
| **Organization** | Prepares R for the appearance of N. |
| **Purpose** | S specifies the purpose of N. |
| **Restatement** | Connects discourse units that are roughly equivalent. |
| **Topic** | S is introduced to steer discourse toward N. |

**# Task Overview**
**Step 1: Extract Stand-Alone Claims**
Decompose the summary into self-contained atomic claims
- No pronouns unless disambiguated.
- Include temporal, causal, and locational context when available.
- Claims should be easy for a human annotator to validate without re-reading the summary.
- Make sure the claims are grouped to the exact amount of original chunk/sentences in the summary input.

**Step 2: Derive Intra-Chunk Relations**
- A **chunk** corresponds to a sentence.
- Use the provided RST tree as guidance (not as absolute ground truth). If the RST tree does not exist in the input, you need to parse the summary following RST.
- Align claims with EDUs where possible.
- Use nucleus/satellite roles and RST relations to construct an intra-chunk relation table.

**Step 3: Derive Inter-Chunk Relations**
- Claims from different sentences may be linked if: The EDUs are connected in the RST tree, **or** You infer temporal, causal, or background links across sentences.
- Potential examples include:
  - Temporal chains
  - Causal progressions
  - Background/support relations

Return a **compact JSON structure** containing:

- A list of claims (‘id‘, ‘text‘, ‘chunk_id‘)
- A list of relations (‘source‘, ‘target‘, ‘rst_label‘, ‘coarse‘, ‘scope‘)

<DEMONSTRATIVE EXAMPLES>

Figure 16: **Full prompt used for Describe claim decomposition.** <DEMONSTRATIVE EXAMPLES> are selected from the GUM dataset, with the corresponding claims constructed through human revisions on LLM-generated decomposition results.

## Structured Inference: Discourse-Based Faithfulness Evaluation Prompt (part 1)

You are an expert in discourse analysis and summarization evaluation. Your task is to assess the *faithfulness* of each summary sentence to its source text using an RST-aware, multi-dimensional framework. Follow the instructions below and think step by step.

**Task Goal**
Evaluate whether each summary sentence is faithful to the **source document** across **literal meaning**, **details**, **discourse structure**, **entity grounding**, and **interpretive stance**.
Faithfulness here means that all key propositions and rhetorical relations are **explicitly present** or **logically entailed** in the source text.

**Input:**

1. Original Document
2. Candidate Summary
3. Summary's discourse-guided claim decomposition and the intra and inter-sentential relation tables.

**Required output format:**
JSON list within the special tag `<FINAL_PRED></FINAL_PRED>`
`[ {"sent_1": [sent_text], "rationale": [rationales collected from thinking process], "faithfulness_label": A}, {"sent_2": [sent_text], "rationale": [rationales ...], "faithfulness_label": B}, ... ]`

**The faithfulness labels should be selected from the following options:**

- **A. Explicitly-Supported:** all atomic facts of the sentence appear verbatim (up to a lexical or syntactic transformation) within the document.
- **B. Generally-Supported:** the document entails the sentence, and the sentence is not explicitly supported by the document. Note that some difference between parts of the sentence and the document is allowed, but it should be restricted to cases in which the part of the sentence adopts a weaker or less certain utterance than the document. If any part of the sentence adopts a stronger or more certain utterance than the document, consider option **D (Fabricated)**.
- **C. Inconsistent:** the sentence logically contradicts the document.
- **D. Fabricated:** the sentence does not logically contradict the document, and is neither logically implied by the document nor by external world knowledge. If any part of the summary sentence adopts a stronger or more certain utterance than the original document, you should assign option **D** (i.e., if the original source said "appears to be" but the summary claim use "is", it would be assigned the label fabricated.)
- **E. Out-Dependent:** the sentence is not logically implied by the document but by the union of the document and external world knowledge.
- **F. Ambiguous:** the sentence or the document has multiple interpretations.
- **G. No-Fact:** the sentence is devoid of facts.

**Step 1 — Align Claims to the Source**
Please carefully examine each sentence by checking the corresponding claims that are produced through some discourse-aware methods.
For each claim that is relevant to this sentence, search the Source Text for sentences or paragraphs that imply the Claim and label:

- **Explicit** – directly stated.
- **Entailed** – implied but logically consistent.
- **Absent / Contradicted** – unsupported or opposite.
- Note if an **attribute**, **spatial location**, or **speaker** is attached to the wrong entity.
- Pay attention to the details.

**Step 2 — Check Relation Fidelity**
Compare how relations appear in the source vs. summary:

1. **Temporal** – event order preserved?
2. **Causal** – causes and effects aligned?
3. **Attribution** – same speaker/thought owner?
4. **Rhetorical** – same function (Cause, Contrast, Elaboration, etc.)?
5. **Perspective** – narrator vs. character unchanged?

This analysis should be performed sentence by sentence. For each sentence, given the assessment of claims that are decomposed from the summary (selecting those relevant to the sentence), you need to validate the intra and inter relations extracted in the relation table versus the original document. For each relation, identify whether the source document implies the relationship described (e.g., if the JSON says "causal" relation between two claims, does the original text use causal language like "because", "due to", or implicitly include the causal relation?)

**Output:** aggregate information relevant for each sentence and organize them with brief reasoning.

Figure 17: Full prompt used for structured inference (1/2).

## Structured Inference (part 2)

**Step 3 — Identify Violation Types**
Leverage the information above to summarize the violation of faithfulness in the table below.

| Category | Definition | Common Errors |
|---|---|---|
| **Factual / Quantitative** | Misstated or fabricated events, numbers, or details | "narrow cave" instead of "narrow shaft" |
| **Entity / Spatial Grounding** | Attribute attached to wrong entity, containment collapsed | Swapping locations or roles |
| **Causal / Temporal** | Event order or cause–effect reversed | "He escaped, then was captured" |
| **Attribution / Perspective** | Wrong speaker, viewpoint, or narrator stance | "He thought" vs. "They thought" |
| **Evaluative / Interpretive** | Added moral or emotional tone | "He was evil"; "A grave threat to humanity" |

Mark each as: None, Minor, or Major.

**Step 4 — Cross-Sentence Relation Validation (CSRE)**

1. Check transitions between adjacent summary sentences ($S_i \rightarrow S_{i+1}$ and $S_i \rightarrow S_{i-1}$) and sentences that are across a longer range. You need to utilize the inter-relations provided.
2. Compare with source texts.
3. Mark **CSRE** type if mismatch:
    - CSRE-Causal
    - CSRE-Temporal
    - CSRE-Contrast
    - CSRE-Attribution
    - CSRE-Evaluation

**Step 5 — Global Summary Final Evaluation**
Providing all the prior information, you are judging the faithfulness of each sentence of a summary to the source document. The faithfulness has seven options:

- **A. Explicitly-Supported:** all atomic facts of the sentence appear verbatim (up to a lexical or syntactic transformation) within the document.
- **B. Generally-Supported:** the document entails the sentence, and the sentence is not explicitly supported by the document. Note that some difference between parts of the sentence and the document is allowed, but it should be restricted to cases in which the part of the sentence adopts a weaker or less certain utterance than the document. If any part of the sentence adopts a stronger or more certain utterance than the document, consider option **D (Fabricated)**.
- **C. Inconsistent:** the sentence logically contradicts the document.
- **D. Fabricated:** the sentence does not logically contradict the document, and is neither logically implied by the document, nor by external world knowledge. If any part of the sentence adopts a stronger or more certain utterance than the document, consider option **D**. If any part of the summary sentence adopts a stronger or more certain utterance than the original document, you should assign option **D** (i.e., if the original source said "appears to be" but the summary claim use "is", it would be assigned the label fabricated.)
- **E. Out-Dependent:** the sentence is not logically implied by the document, but by the union of the document and the external world knowledge.
- **F. Ambiguous:** the sentence or the document has multiple interpretations.
- **G. No-Fact:** the sentence is devoid of facts.

While assessing **B** versus **D**, you should allow for **Contextual compression**, **Metonymy** (place ↔ residence ↔ area) and **Background consolidation**, and do not separate fact from neighboring context too strictly.

The final output would be a JSON list within the special tag <FINAL_PRED></FINAL_PRED>: `[ {"sent_1": [sent_text], "rationale": [rationales collected from thinking process], "faithfulness_label": A}, {"sent_2": [sent_text], "rationale": [rationales ...], "faithfulness_label": B}, ... ]`

**### Document:**
DOCUMENT_PLACEHOLDER

**### Summary with N sentences:**
SUMMARY_PLACEHOLDER

**### Summary Claims and Relation Table:**
CLAIMS_PLACEHOLDER

Figure 18: Full prompt used for structured inference (2/2).