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 *Equal senior contribution

Motivation

- Errors in long-form LLM summaries (**≥4 sentences**) are subtle and context-dependent, not surface-level.
- **Decompose-then-verify** splits summaries into atomic claims, but atomicity drops the **discourse relations** (cause, contrast, attribution, …) that bind claims together.
- Fact-checkers lean on surface semantic matching and miss relation-level errors.

Source Document

… I remember Mom, a sweet woman. Her smile, her strong hands, and **witty humor** … After several minutes, I gather the courage to speak to her … Mom's voice answers me in return: **You brat! I took care of you and your brother! I gave you your own closet. But you scratched it like a cat! You! You didn't know how to wash in the tub, so I showed you how!** I laugh at the whisper. "You're telling a different story, mom." I smile. **"At least you won't hurt me or Hector anymore."**

Key evidence: Mother's actual dialogue is scolding, not witty/humorous.

Summary

The narrator is visiting a cemetery, … they come across several graves that …
The narrator talks to their mother's grave and reminisces, **and finds that even in death, their mother still has plenty of wit and humor.**

Human Annotation

Label: Unfaithful
Rationale: The mother was nasty and cruel which is not captured by her being full of "wit and humor".

How do evaluation frameworks assess this sentence?

Prior Work ✗ Incorrect

Prediction: Faithful

Rationale: [Extraction] The narrator reminisces: *"I remember Mom, a sweet woman … and **witty humor**"* … [Inference] The narrator's laughter at the mother's posthumous remarks, combined with recalling her "witty humor" supports that the mother still displays humor in death. [Conclusion] So, the statement is [Attributable]

⚠ *Relies on surface matching of "witty humor" without checking whether the mother's dialogue supports it*

This Work (Describe) ✓ correct

Prediction: Unfaithful

Discourse-based Decomposition:
C7: The narrator talks to the narrator's mother's grave.
C8: The narrator reminisces while at the narrator's mother's grave.
C9: The narrator finds that the deceased mother still displays wit and humor.
Relations: (C8, C7, Elaboration), (C7, C9, **Mode-Means**)

Rationale: … Unsupported/fabricated: "finds that even in death, … and humor" is not in the text. The mother's actual postmortem speech is scolding *("You brat!...")*, and the narrator says, "At least you won't hurt me or Hector anymore," which does not display wit/humor.
Relation fidelity issue: The implied Mode/Means that that talking to her reveals wit/humor is absent; attribution conflates narrator's memory (living mom's humor) with the dead mother's current voice.

*Prior work matches "witty humor" on the surface and says faithful; **DESCRIBE** checks the Mode–Means relation against the source and flags it.*

Our Approach: DESCRIBE

Source Doc

… 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. The Colombian followed Nicklaus with an ace on the fourth and also managed one on the eighth , **but he couldn't quite land the title, losing out to world no.74 Kevin Streelman in a play-off.** …

Summary (w/ annotation)

S1: Jack Nicklaus, at 75 years old, hit an ace during the par-3 contest ..
S2: He achieved this on the 130-yard fourth hole, impressing the crowd …
S3: This was Nicklaus's first hole-in-one at Augusta, though Camilo Villegas matched the feat on the same hole, along with another on the eighth, **leading to a play-off where Kevin Streelman won.**

Human Annotation

S3 contains fabricated error.
"note": "Weak causal relationship", "rationale": "he couldn't quite land the title, losing out to world no 74 kevin streelman in a play-off"

Prior Work

S3 has no error. The transcript confirms it was his first at Augusta, Villegas had aces on the fourth and eighth, and Streelman won in a play-off.

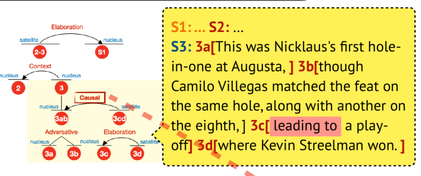
Relied on surface-level text matching

Our Work: *Describe*

Part 1: Discourse-guided Claim Extraction (Section 4.2)

Step 1: Discourse Parsing

RST Parser



S1: … S2: …
S3: 3a[This was Nicklaus's first hole-in-one at Augusta,] 3b[though Camilo Villegas matched the feat on the same hole, along with another on the eighth,] 3c[leading to a play-off] 3d[where Kevin Streelman won.]

Summary RST parse tree

Step 2: Claim Decomposition

Claims

[{ C1: Jack Nicklaus hit a hole-in-one during the par-3 contest at Augusta National. src:S1 }, …
{ C6: The hole-in-one at the par-3 contest was Jack Nicklaus's first hole-in-one at Augusta National. Src: S3}
{ C7: Camilo Villegas hit a hole-in-one on the fourth hole during the contest, src: S3},
{ C8: Camilo Villegas hit a hole-in-one on the eighth hole during the contest. Src: S3},
{ C9: Kevin Streelman won a play-off at the par-3 contest at Augusta National. Src: S3}, …]

Relations

[… (C6, C7, Adversative-Concession); (C7, C9, Causal-Result), (C8, C9, Causal-Result), …]

Part 2: Structured Inference (Section 4.3)

Doc Summary Describe claims (Claims Relations)

For each summary sent
(1) claim alignment and verification with source
(2) intra-sentence discourse relation validation
(3) cross-sentence relation checking

Describe Output

Sent_3: This was Nicklaus's first hole-in-one at Augusta, though Camilo …
Rationale: { Claim alignment: C6 (first HIO at Augusta) Explicit; C7 (Villegas ace on the fourth) Explicit; C8 (Villegas ace on the eighth) Explicit; C9 (Streelman won a play-off) Explicit.
Relation fidelity: The summary's **internal CAUSAL link ('leading to a play-off')** between Villegas's aces and the play-off outcome (C7/C8→C9) is not explicitly supported; source only states Villegas lost to Streelman in a play-off. Treat as temporal/culmination compression rather than proven causation.
Violations: Minor Causal/Temporal inflation (implies causality not stated)
Cross-Sentence-Relation-Evaluation: S2→S3 Contrast/Concession is consistent with source narrative (Nicklaus's feat then others' feats eclipsing it).' }
Predicted Label: Fabricated

1 Discourse Parsing

DMRST retrained on **GUM v11** (16 genres, 15 relation classes) + DisCut EDU segmentation produce an RST tree of the summary.

2 Claim Decomposition

LLM writes self-contained atomic claims plus a **directed relation graph**: intra- and inter-sentential RST relations between claims.

3 Structured Inference

For each sentence: **align** claims to source → **validate** intra-sentence relations → **check** cross-sentence relations → one of 7 labels + rationale.

Research Questions

RQ1 Do existing decompositions preserve discourse in long summaries?

RQ2 Do discourse-guided claims + structured inference improve evaluation?

RQ3 Which components drive the gains, and for which types of sentences?

RQ4 Do rationales align with human reasoning and discourse?

Benchmarks & Baselines

Dataset	Domain	Summ. / Sent.	Labels
STORYSUMM[1]	Narratives	96 / 580	Binary
VERIGRAY[2]	News	412 / 2,044	7-way fine-grained

- **Fine-tuned:** MiniCheck, ClearCheck, FaithLens
- **Discourse-inspired:** StructScore, FDSpotter
- **Prompting:** Fewshot, FineSURE, VG-prompt on Qwen3-30B, gpt-oss-120B, Gemini-3-Flash, GPT-5
- **Metrics:** Balanced Accuracy, MCC, hallucination P/R/F1

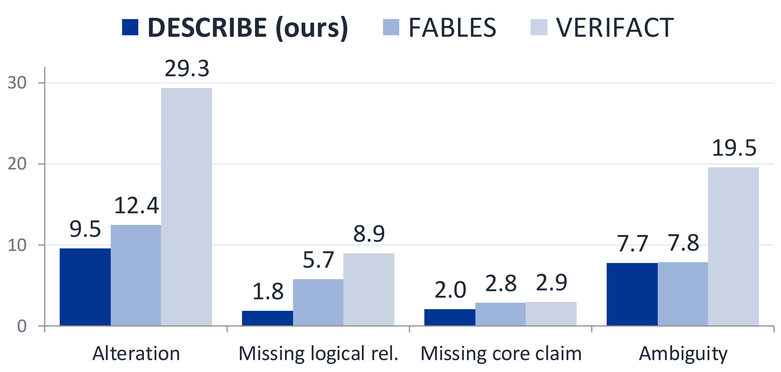
Takeaways

- Atomic claims are not enough: **verify the relations** that connect them.
- RST-guided claims + structured inference give **consistent gains** on fine-grained labels.
- Discourse-driven rationales offer higher quality, yet they may not correlate with sentence-level label accuracy. Future benchmarks should annotate **relation-level errors** below the sentence level.

References
[1] Melanie Subbiah, Faisal Ladhak, Akankshya Mishra, Griffin Thomas Adams, Lydia Chilton, and Kathleen McKeown.STORYSUMM: Evaluating Faithfulness in Story Summarization, EMNLP 2024
[2] Ding, Qiang, et al. "The gray zone of faithfulness: Taming ambiguity in unfaithfulness detection." *arXiv preprint arXiv:2510.21118* (2025).
[3] 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)*, 2026

RQ1: Decomposition Errors

Decomposition errors by type (% of outputs)



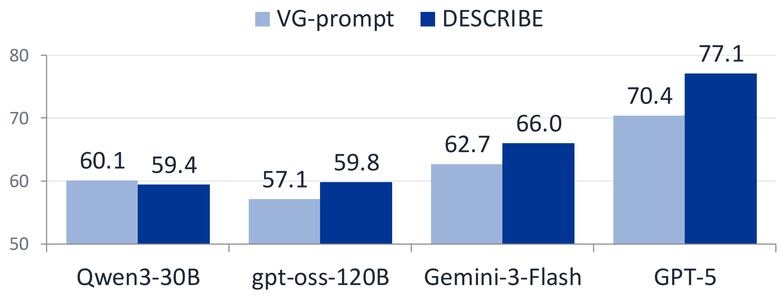
Both benchmarks combined; judged with Hu et al. (2025) prompts (92% agreement with manual review). Over-decomposition not shown.

RQ2: Main Results

Method	STORYSUMM BAcc		VERIGRAY (sentence)		
	Summ.	Sent.	BAcc	MCC	Hal. F1
Fine-tuned fact-checkers					
MiniCheck	53.6	58.1	59.6	0.23	32.4
FaithLens	54.7	57.7	59.1	0.27	31.1
Discourse-inspired					
StructScore	61.1	67.5	64.4	0.26	39.6
FDSpotter	64.2	58.6	54.5	0.15	19.3
Prompting, GPT-5 backbone					
Fewshot	62.4	72.0	81.3	0.66	71.0
FineSURE	66.4	**75.0**	81.6	0.66	70.9
VG-prompt	66.9	72.9	82.5	0.64	70.4
DESCRIBE	**69.4**	71.4	**84.9***	**0.73***	**77.1***

*Partial Table 2. Bold = best in column; * = significantly better than VG-prompt (paired bootstrap, $p < 0.05$).*

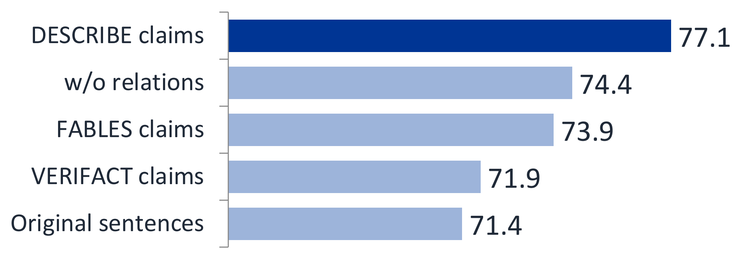
VERIGRAY hallucination detection F1 across backbones



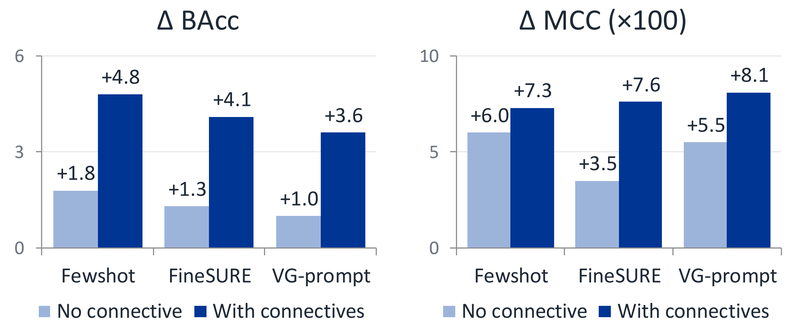
VERIGRAY: DESCRIBE beats the strongest baseline on BAcc for every backbone (≈2 points); F1 gains grow with backbone strength, up to **+6.7 F1** with GPT-5.
STORYSUMM: mixed at sentence level, as its unanimity-based binary labels hide the relational errors DESCRIBE targets [3].

RQ3: What Drives the Gains?

Same structured inference, different claims (VERIGRAY Hal. F1)



DESCRIBE claims alone also lift plain VG-prompt inference compared to original sentences (73.8 vs. 70.4 F1).



Sentences **with discourse connectives** obtain larger Gains (MCC +8.1 vs. +5.5 over VG-prompt) → Complex sentences benefit more.

RQ4: Trustworthy Rationales

System	Completeness	Human Rationale Alignment	Discourse
DESCRIBE	**95.2**	**92.7**	**92.1**
VG-prompt	88.5	90.5	82.4
FineSURE	70.4	85.2	74.8
FaithLens	92.4	72.0	58.5

Result on VERIGRAY, rationales are analyzed over correct predictions using Gemini-3-Flash as a judge

Even on wrong predictions: DESCRIBE better matches human rationales (39.8% vs. 30.9%) and preserves more discourse structure (83.7% vs. 64.7%) than VG-prompt.

