# Discourse-Driven Evaluation: Unveiling Factual Inconsistency in Long Document Summarization

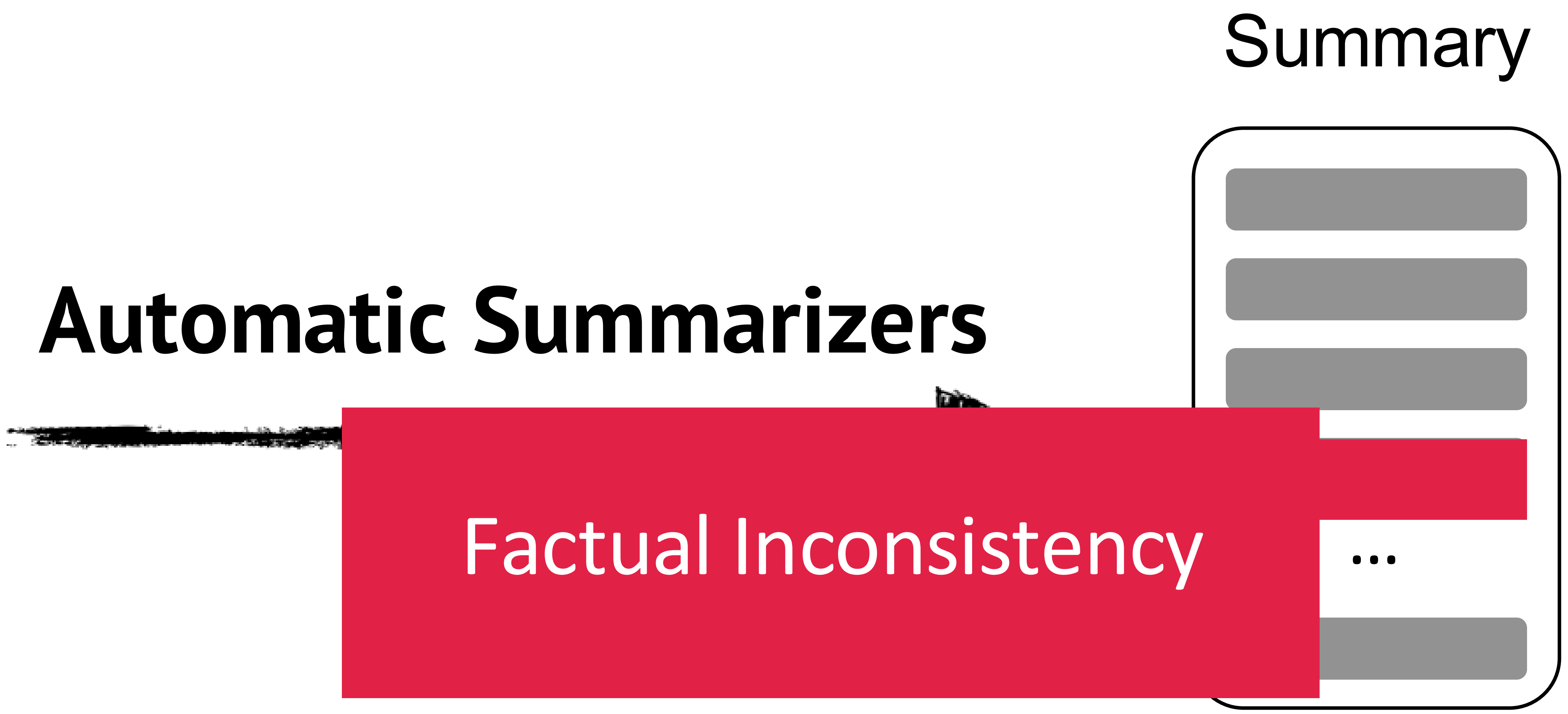
**Yang Zhong** and Diane Litman

NAACL 2025

University of Pittsburgh

# Factual Inconsistency Evaluation

Document >= 2000 words



Summary

Automatic Summarizers

Factual Inconsistency

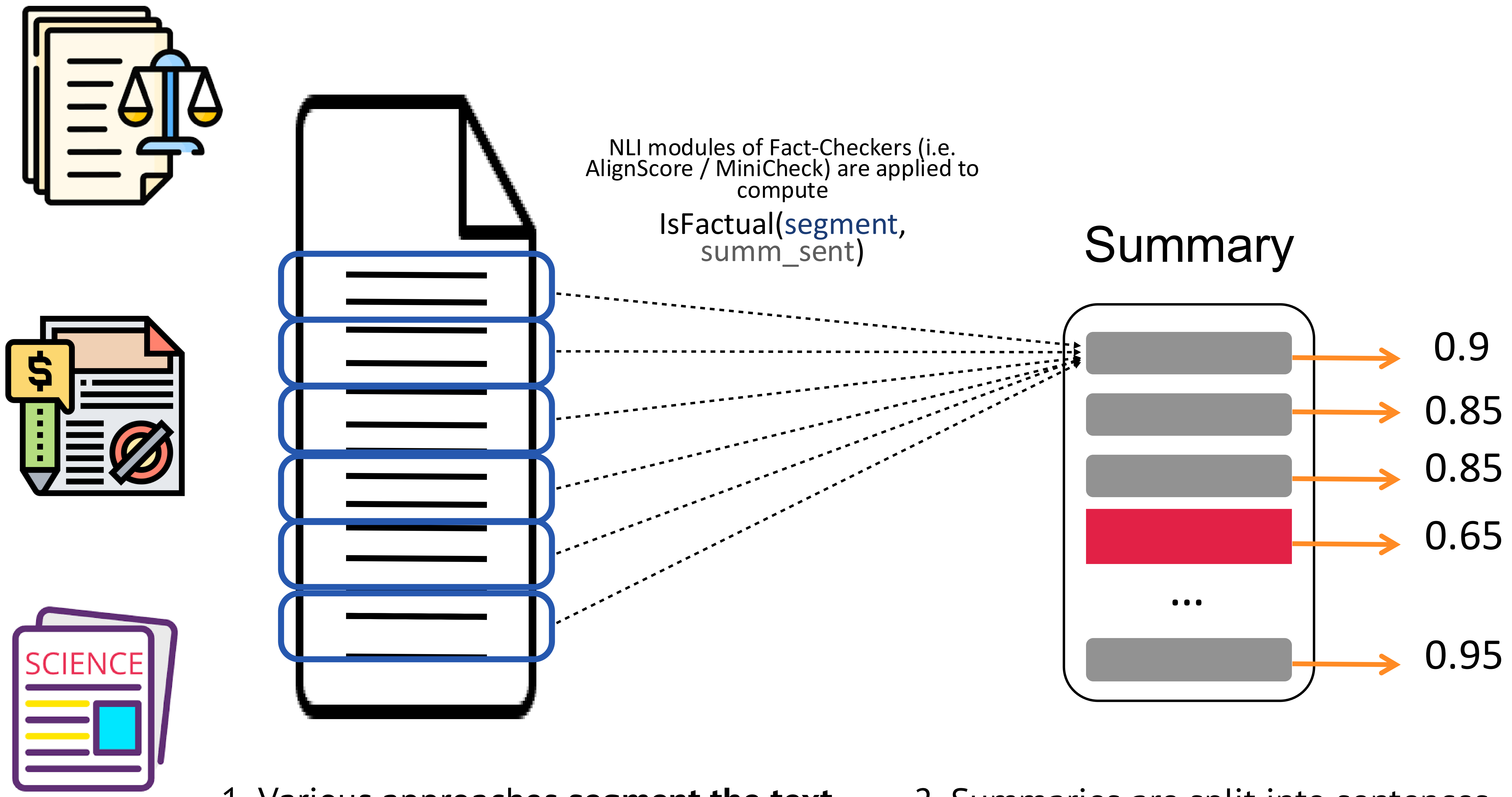
…

However, summaries can contain incorrect information which

a. Does not appear in the source document

b. Can not be inferred  from the source document

# Factual Inconsistency Evaluation


SCIENCE
$
NLI modules of Fact-Checkers (i.e. AlignScore / MiniCheck) are applied to compute
IsFactual(segment, summ_sent)
Summary
0.9
0.85
0.85
0.65
...
0.95

Mean: ~0.85

Min: 0.65

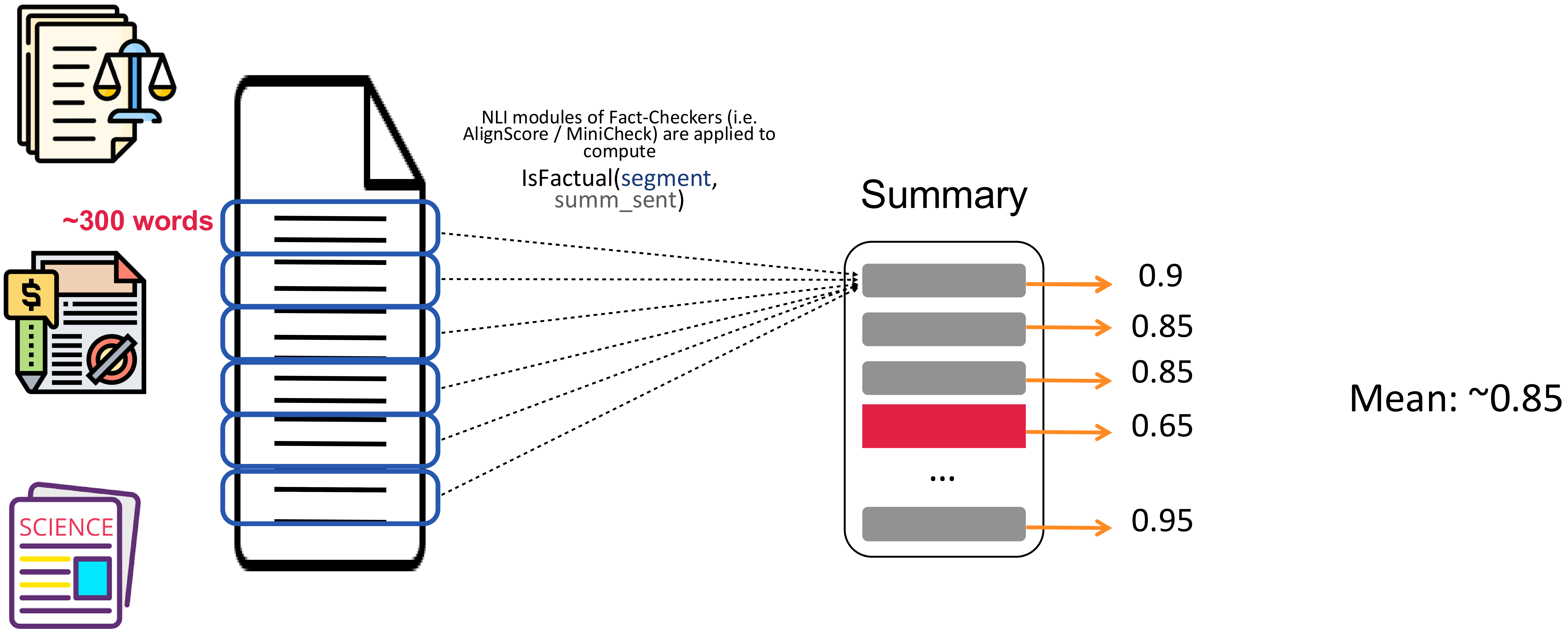
1. Various approaches **segment the text** into smaller units, such as continuous chunks or even individual sentences.

2. Summaries are split into sentences, then evaluated individually.

3. The summary-level score is aggregated by averaging the sent-level scores or select the minimum.

Detecting factual inconsistency for long document summarization remains challenging!

# Factual Inconsistency Eval Challenges



~300 words

NLI modules of Fact-Checkers (i.e. AlignScore / MiniCheck) are applied to compute

IsFactual(segment, summ_sent)

Summary

Mean: ~0.85

**Challenge 1**

Article structure may not be well preserved when using continuous chunking.

**Challenge 2**

Taking an unweighted average can hide mistakes in individual sentences, leading to inaccurate evaluation

# Our Work

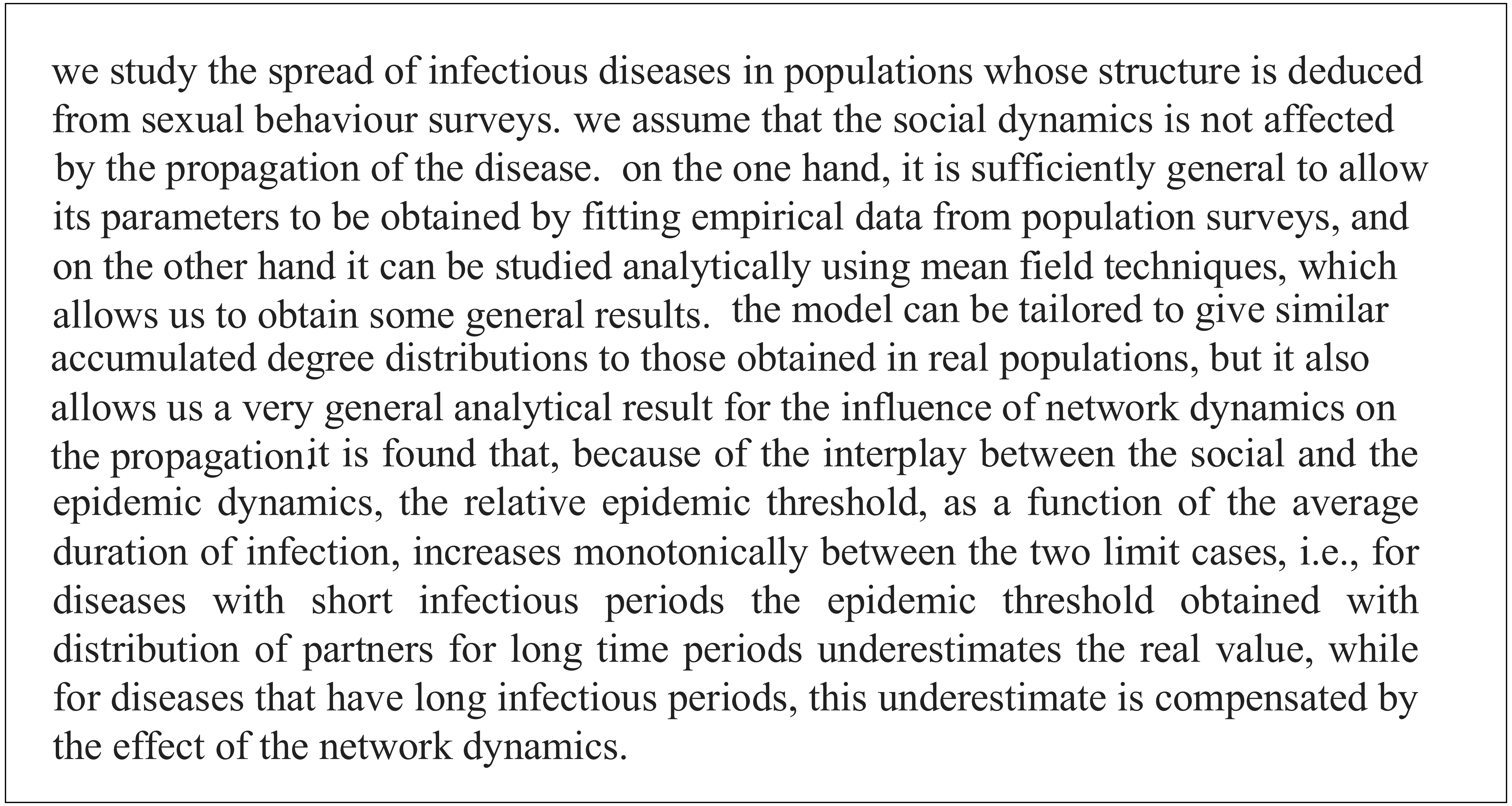
- Analysis of discourse level factors related to the factual inconsistency
  - Discourse Analysis on Summary Errors
  - Document Structure
- Using linguistic features to enhance summary-level factual consistency evaluation

# Our Work

- Analysis of discourse level factors related to the factual inconsistency
  - Discourse Analysis on Summary Errors
  - Document Structure
- Using linguistic features to enhance summary-level factual consistency evaluation

# Discourse Analysis

Below is one example of machine-generated summary of an arXiv paper



we study the spread of infectious diseases in populations whose structure is deduced from sexual behaviour surveys. we assume that the social dynamics is not affected by the propagation of the disease. on the one hand, it is sufficiently general to allow its parameters to be obtained by fitting empirical data from population surveys, and on the other hand it can be studied analytically using mean field techniques, which allows us to obtain some general results. the model can be tailored to give similar accumulated degree distributions to those obtained in real populations, but it also allows us a very general analytical result for the influence of network dynamics on the propagation. it is found that, because of the interplay between the social and the epidemic dynamics, the relative epidemic threshold, as a function of the average duration of infection, increases monotonically between the two limit cases, i.e., for diseases with short infectious periods the epidemic threshold obtained with distribution of partners for long time periods underestimates the real value, while for diseases that have long infectious periods, this underestimate is compensated by the effect of the network dynamics.

# Factual Consistency

## And Discourse Analysis

[1 We study the spread of infectious diseases in populations whose structure is deduced from sexual behaviour surveys.]

[2 We assume that the social dynamics is not affected by the propagation of the disease. ]

[3a On the one hand, it is sufficiently general] [3b to allow its parameters to be obtained by fitting empirical data from population surveys,] [3c and on the other hand it can be studied analytically using mean field techniques,] [3d which allows us to obtain some general results.

[4a The model can be tailored] [4b to give similar accumulated degree distributions to those obtained in real populations,] [4c but it also allows us a very general analytical result for the influence of network dynamics on the propagation. ]

[5a It is found that, because of the interplay between the social and the epidemic dynamics, the relative epidemic threshold, as a function of the average duration of infection, increases monotonically between the two limit cases,] [5b for diseases with short infectious periods the epidemic threshold obtained with distribution of partners for long time periods underestimates the real value,] [5c while for diseases that have long infectious periods, this underestimate is compensated by the effect of the network dynamics.]

No error

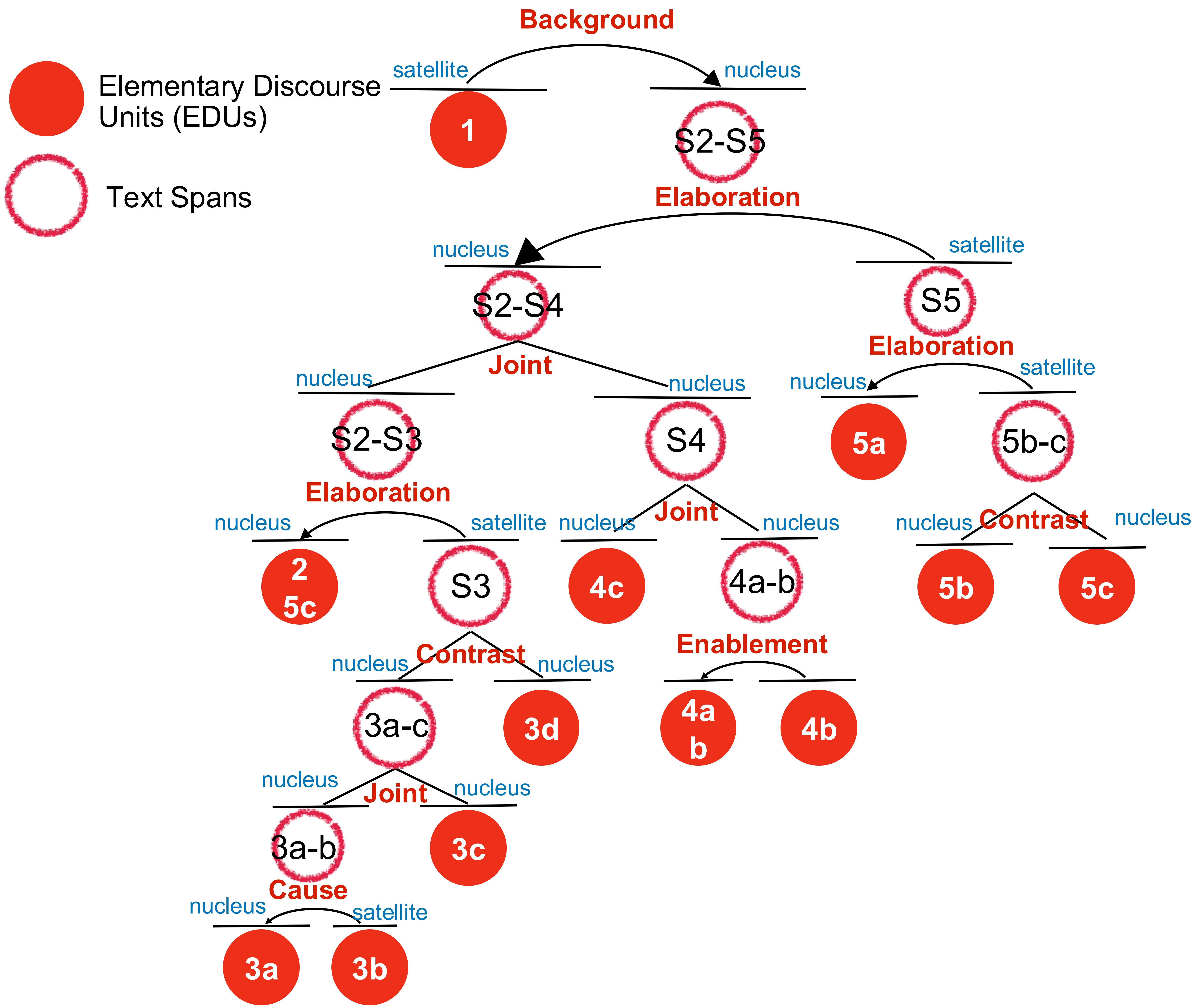
No error

No error

No error

Linkage error

# Rhetorical Structure Theory (RST)

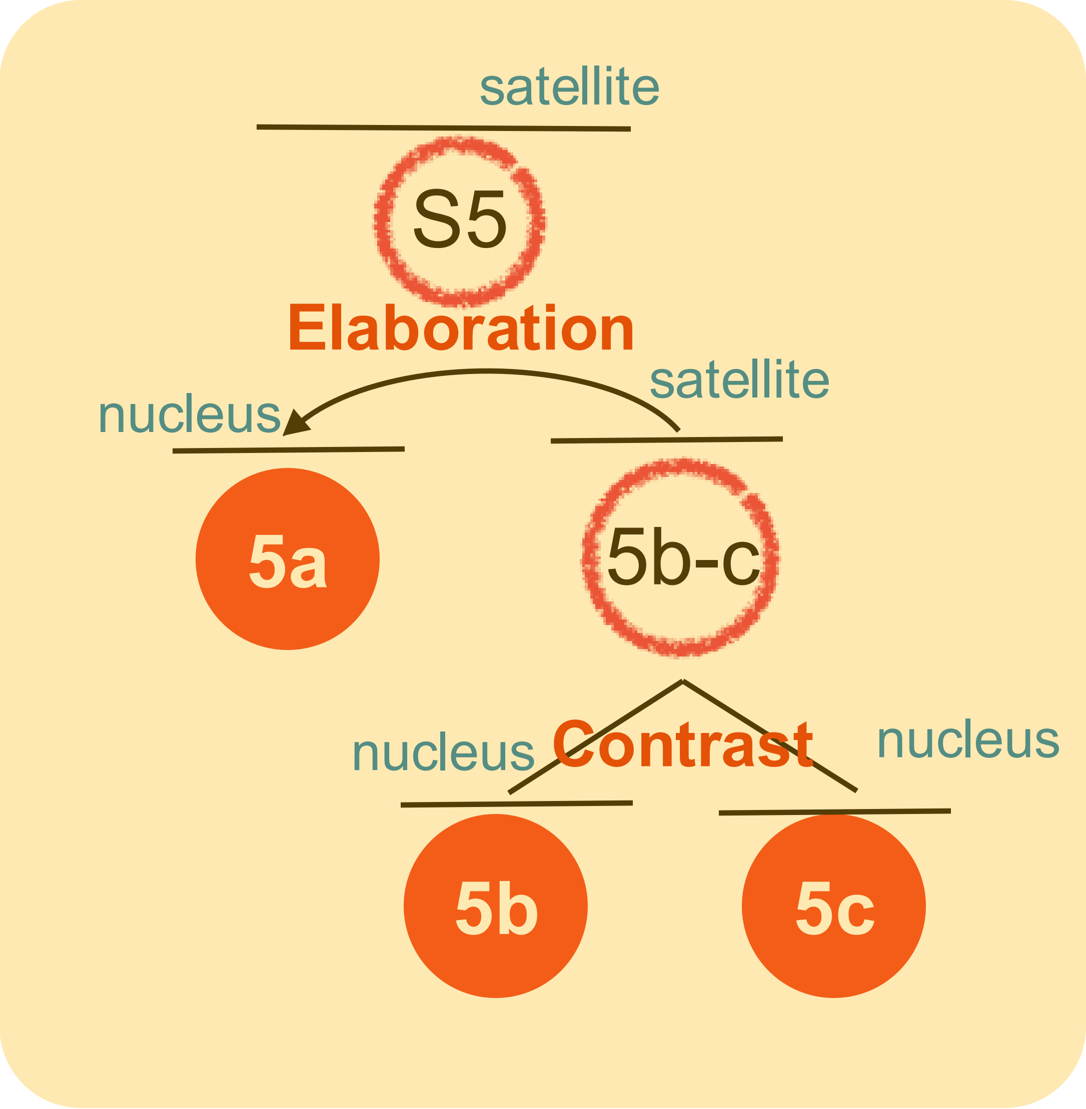


* using the DMRST discourse parser from Liu et al. (2021)

# Rhetorical Structure Theory (RST)

● Elementary Discourse Units (EDUs)

○ Text Spans



5a it is found that, because of the interplay between the social and the epidemic dynamics, the relative epidemic threshold, as a function of the average duration of infection, increases monotonically between the two limit cases,

5b i.e., for diseases with short infectious periods the epidemic threshold obtained with distribution of partners for long time periods underestimates the real value,

5c while for diseases that have long infectious periods, this underestimate is compensated by the effect of the network dynamics.

* using the DMRST discourse parser from Liu et al. (2021)

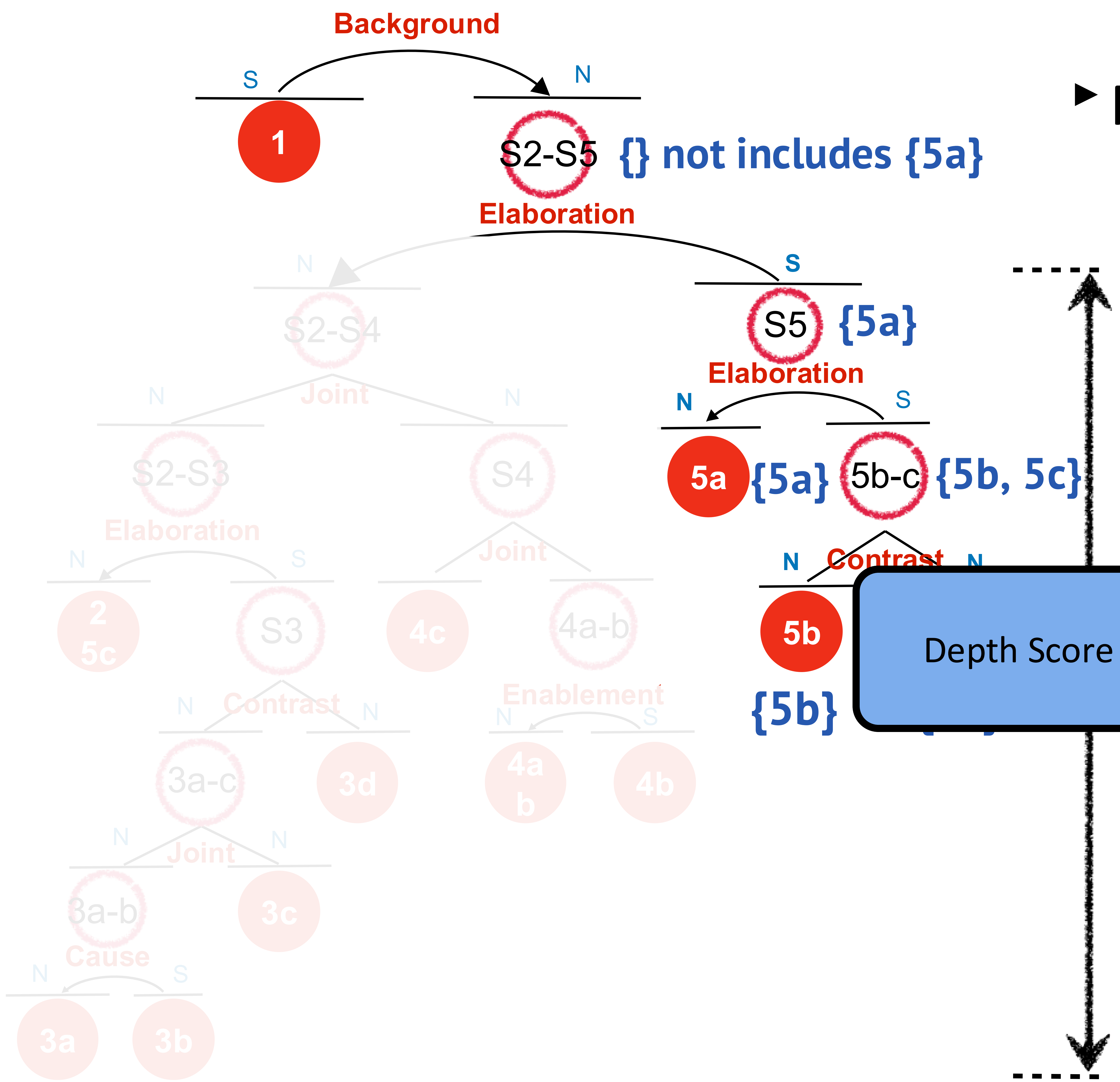
# Motivation

Previous studies show that selecting salient nucleus sentences can enhance summarization performance. [1, 2]

Our work takes a different direction by exploring the relationship between discourse features and factual consistency evaluation.

[1]  Annie Louis, Aravind Joshi, and Ani Nenkova. *Discourse indicators for content selection in summarization.* SIGDIAL 2010
[2] Dongqi Liu, Yifan Wang, and Vera Demberg *Incorporating Distributions of Discourse Structure for Long Document Abstractive Summarization*. ACL 2023

# Explored Discourse Features



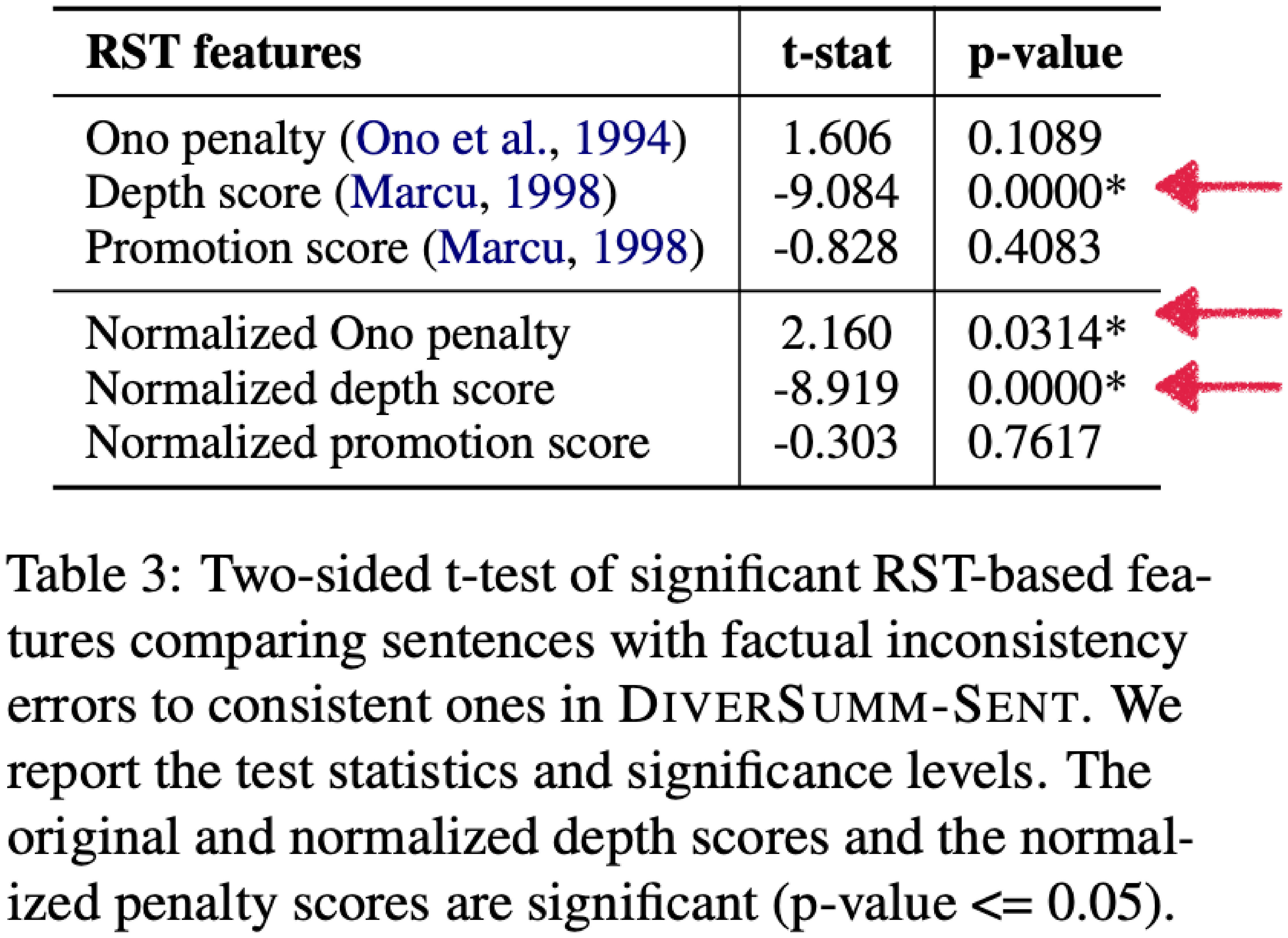
- Promotion Depth Score [1]
  - Motivation: rewarding nucleus status by recording a prompt set {} for each node.
  - Hypothesis: units in the promotion sets of nodes close to the root are hypothesized to be more tant
  - Design: The depth of the tree from the highest promotion is assigned as the score for that EDU

Depth Score of S5 = 5

[1] Daniel Marcu. 1998. To build text summaries of high quality, nuclearity is not sufficient.

# Promotion Depth Score

- Compare “non-factual” sentences with “factual” sentences



| RST features | t-stat | p-value |
|---|---|---|
| Ono penalty (Ono et al., 1994) | 1.606 | 0.1089 |
| Depth score (Marcu, 1998) | -9.084 | 0.0000* |
| Promotion score (Marcu, 1998) | -0.828 | 0.4083 |
| Normalized Ono penalty | 2.160 | 0.0314* |
| Normalized depth score | -8.919 | 0.0000* |
| Normalized promotion score | -0.303 | 0.7617 |

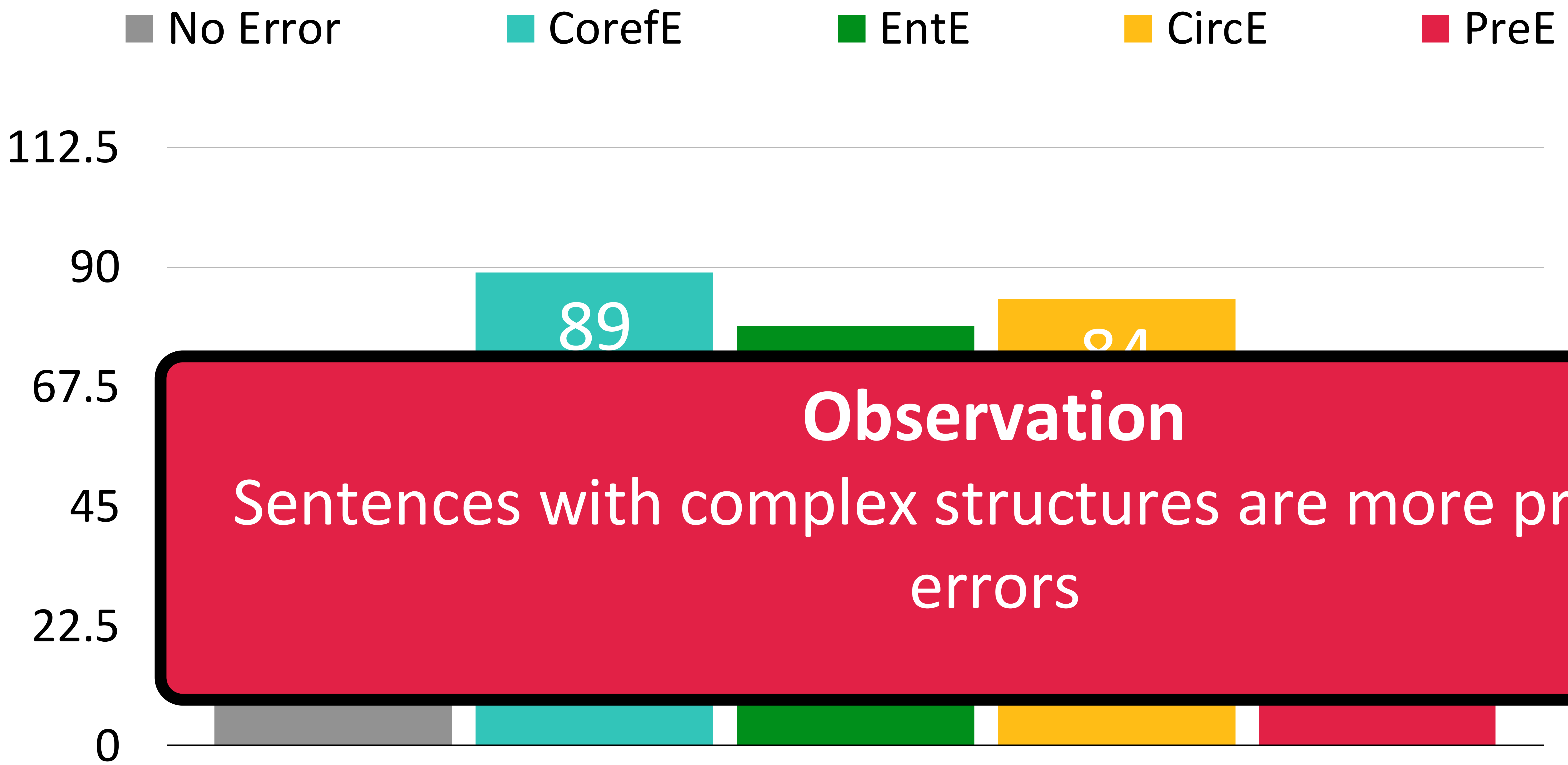
Table 3: Two-sided t-test of significant RST-based features comparing sentences with factual inconsistency errors to consistent ones in DIVERSUMM-SENT. We report the test statistics and significance levels. The original and normalized depth scores and the normalized penalty scores are significant (p-value <= 0.05).

**Observation**

Errors are associated with the nuclearity and discourse feature

# Complexity of the Sentence

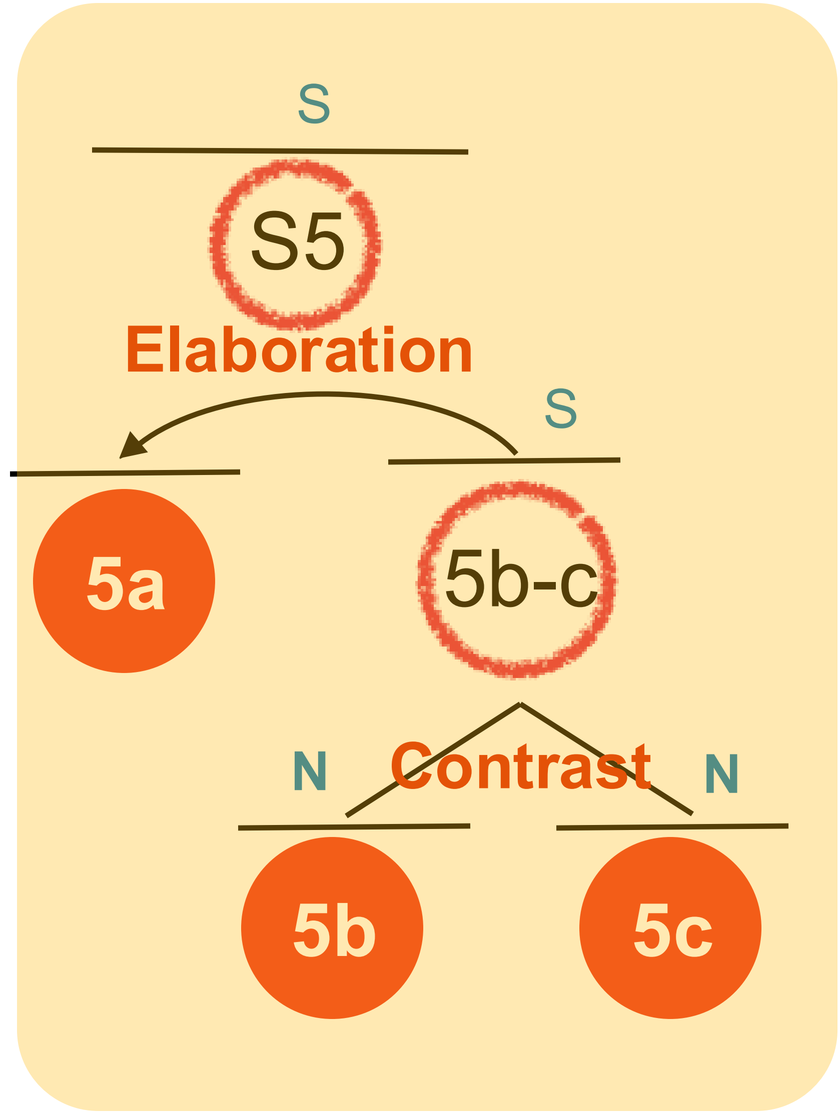
- We evaluate the distribution of discourse-subtree depths for sentences.



Percentage of NoError and factually inconsistent sentences with dense structure (depth >= 2)

**Observation**

Sentences with complex structures are more prone to errors

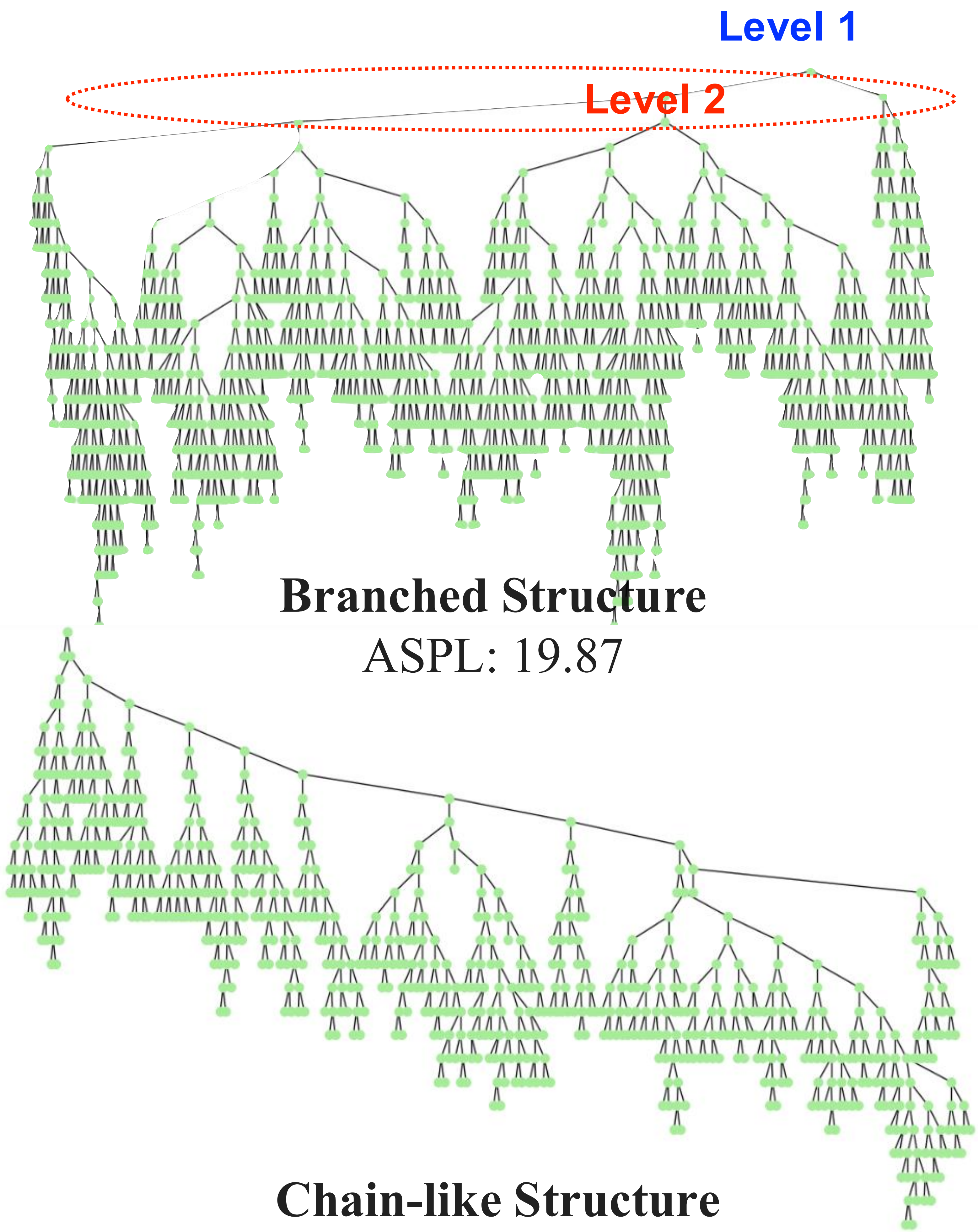


# Our Work

- Analysis of discourse level factors related to the factual inconsistency
  - Discourse Analysis on Summary Errors
  - Document Structure
- Using linguistic features to enhance summary-level factual consistency evaluation

# Discourse Structure Inspired Segmentation

- Through RST parsing, we observe long documents exhibit varying structures



**Branched Structure**
ASPL: 19.87

**Chain-like Structure**
ASPL: 17.02

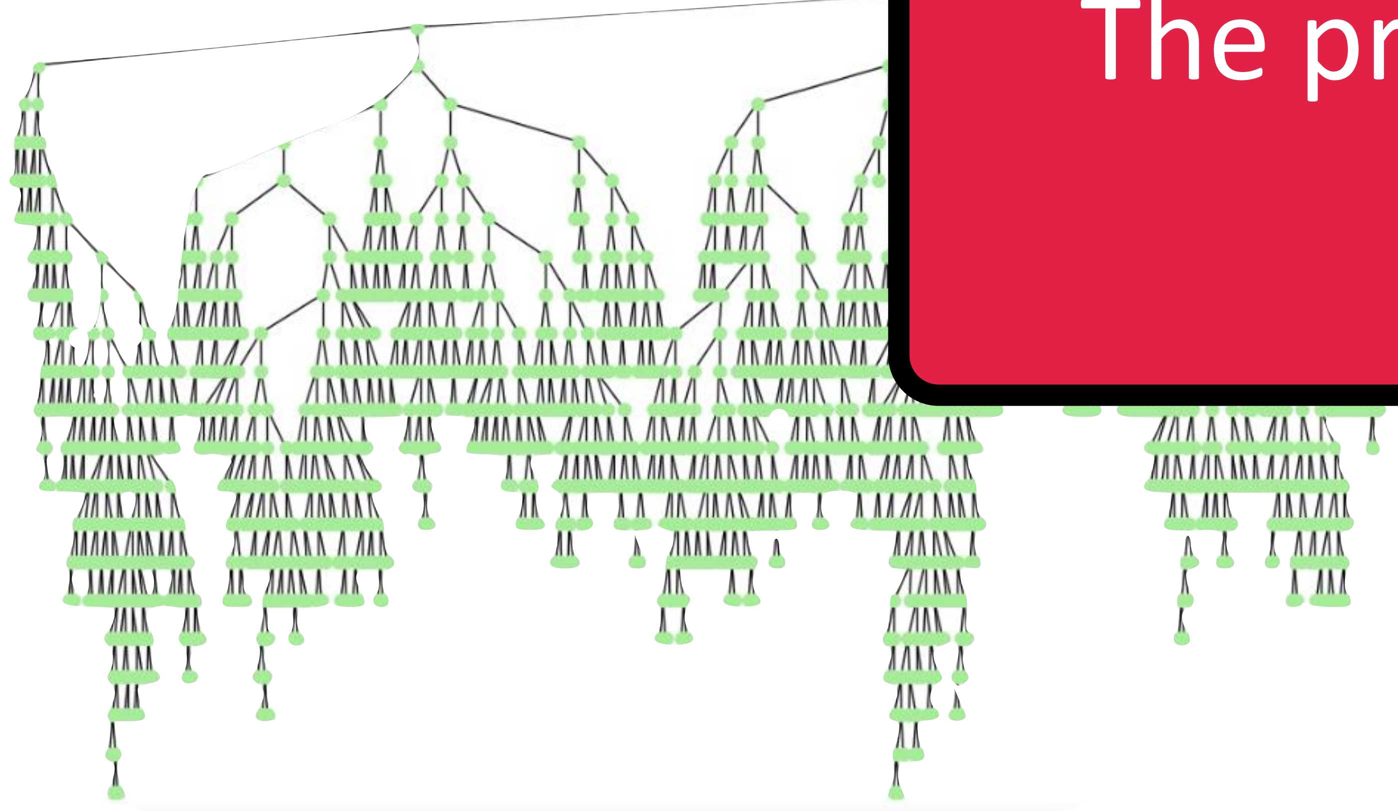
# Discourse Structure Inspired Segmentation

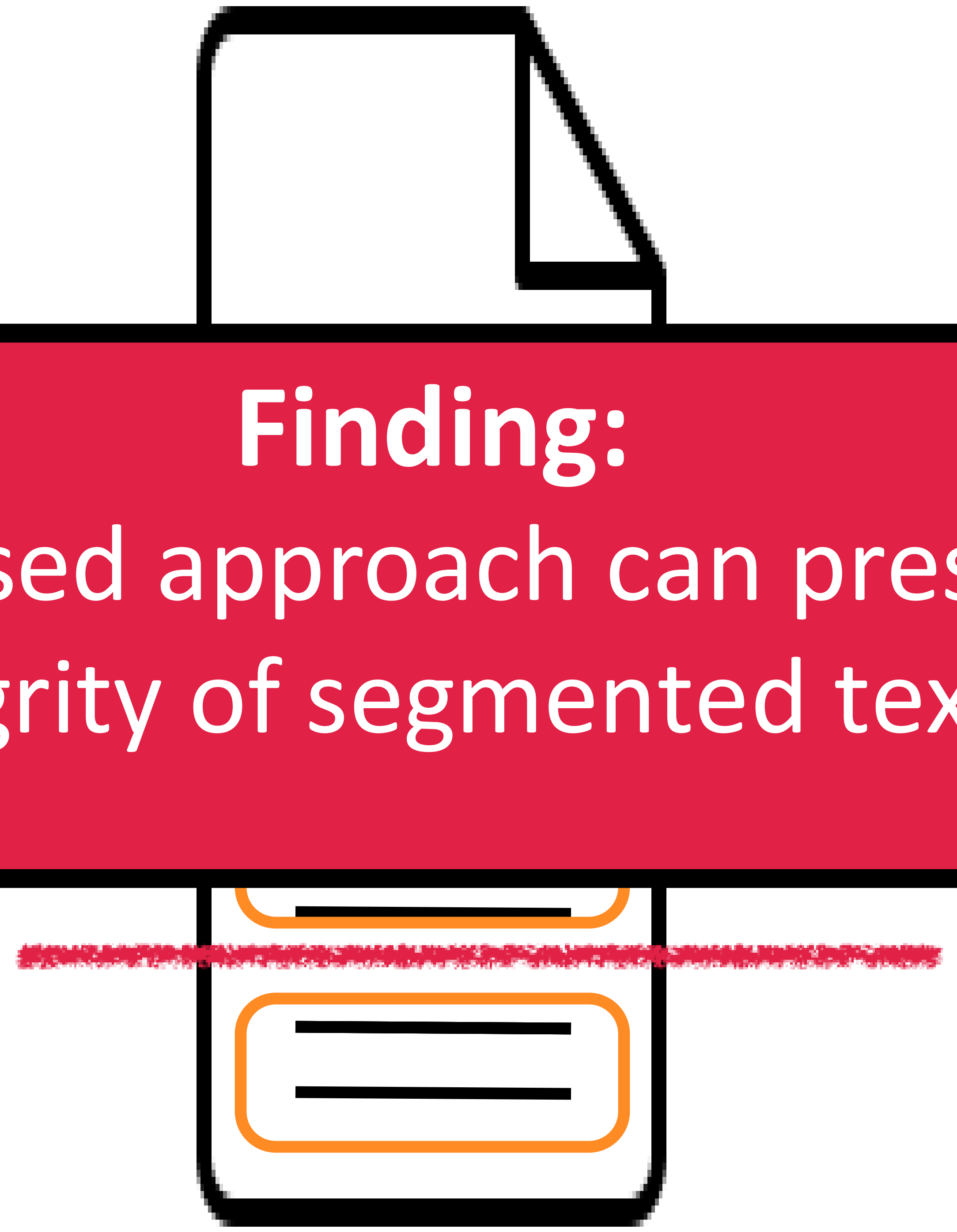
- We propose incorporating the high level discourse inspired structures, and further preserve the document structures according to the document hierarchies

Level 1

Segment 1: EDUs [1-64

Segment 2: EDUs [649-6





**Finding:**
The proposed approach can preserve the integrity of segmented texts.

**Step 1 Segment:** break down the nto two parts

**ment:** Split the according to its hierarchical structure:

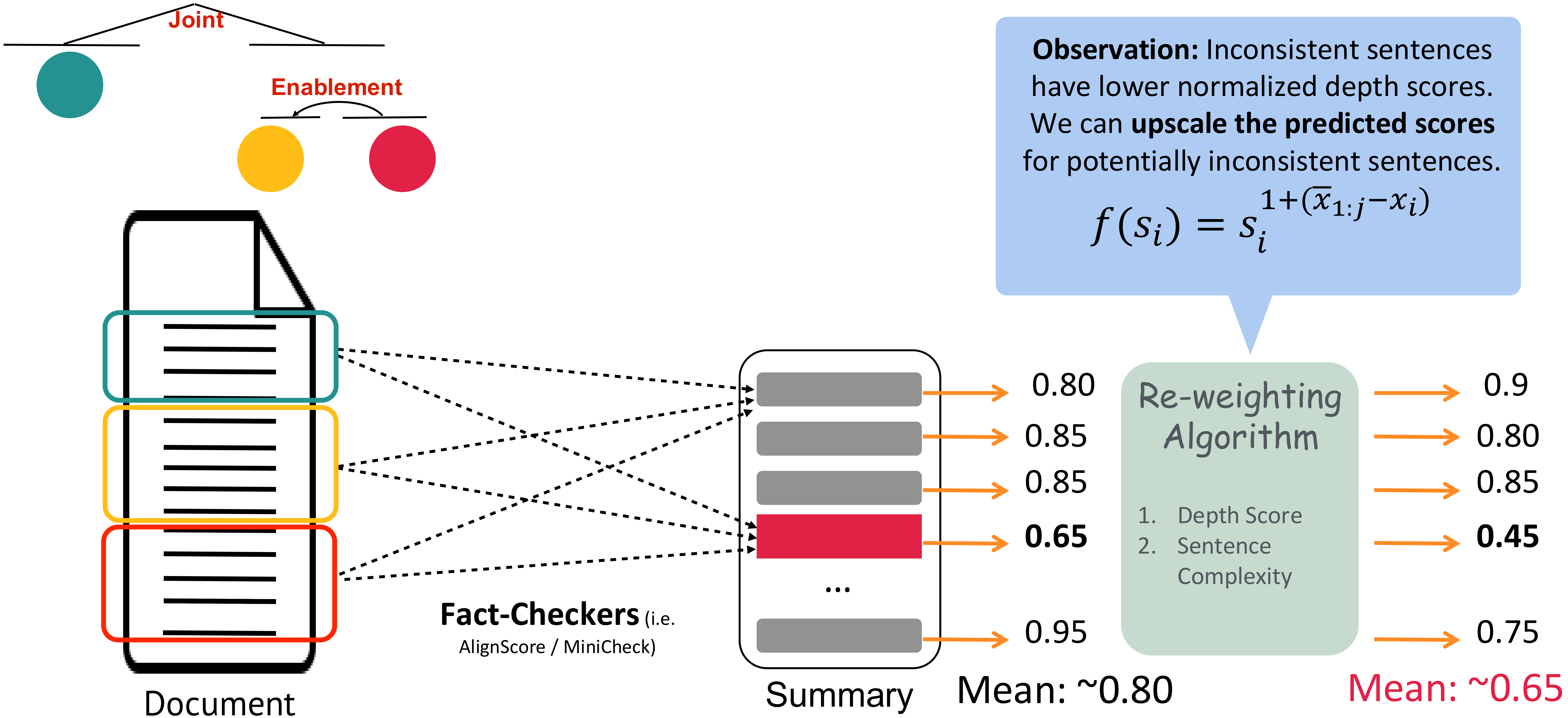
**Paragraphs**

**Sentences**

# Our Work

- Analysis of discourse level factors related to the factual inconsistency
  - Discourse Analysis on Summary Errors
  - Document Structure
- Using linguistic features to enhance summary-level factual consistency evaluation

# Our Approach — StructScore



**Observation:** Inconsistent sentences have lower normalized depth scores. We can **upscale the predicted scores** for potentially inconsistent sentences.

$$f(s_i) = s_i^{1+(\overline{x}_{1:j}-x_i)}$$

**Contribution 1:**
Discourse-Structure-Inspired Source Segmentation

**Contribution 2:**
Sentence Reweighting Based on Discourse Features

# Experimental Setup

- Datasets

  Multiple long document summarization evaluation datasets which cover diverse domains:
  **DiverSumm** (ArXiv, GovReport, ChemSum .. etc), **LegalSumm** (legal)
  **LongEval** (Pubmed) ...

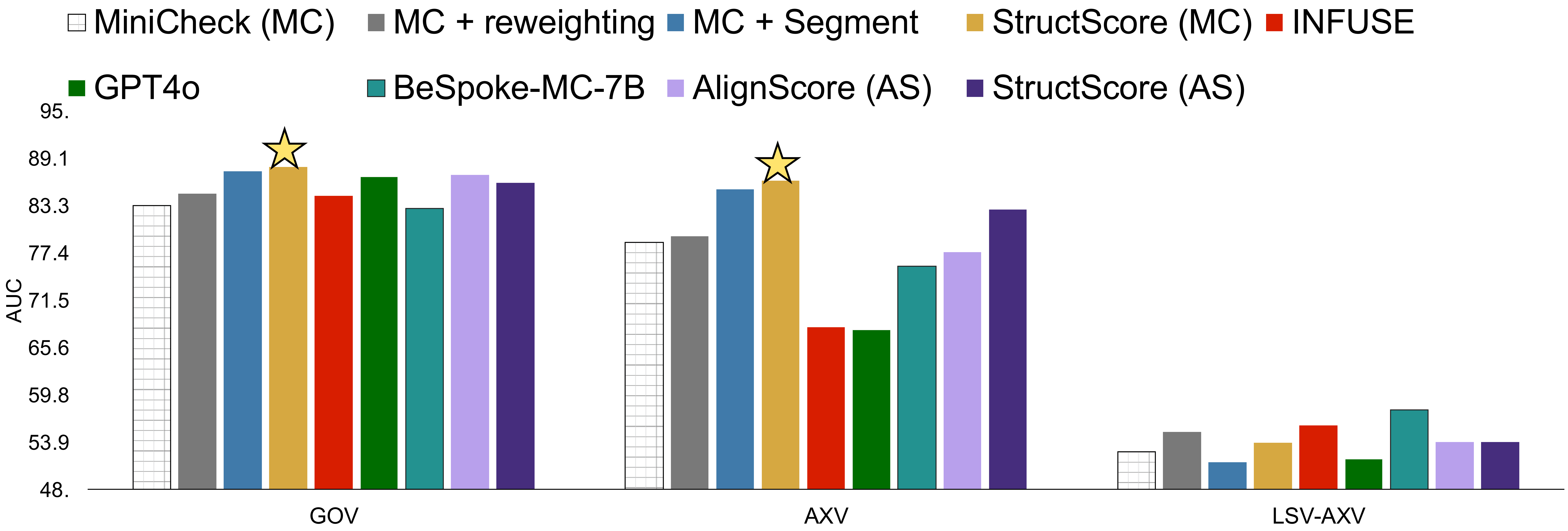
- Baselines
  - Long summary evaluation specialized models: **INFUSE** and LongDocFactScore
  - LLM-based models: **GPT4o** and **BeSpoke-MC-7B**
  - Strong NLI-based models with limited context: **AlignScore** and **MiniCheck**

+

StructScore

# Results

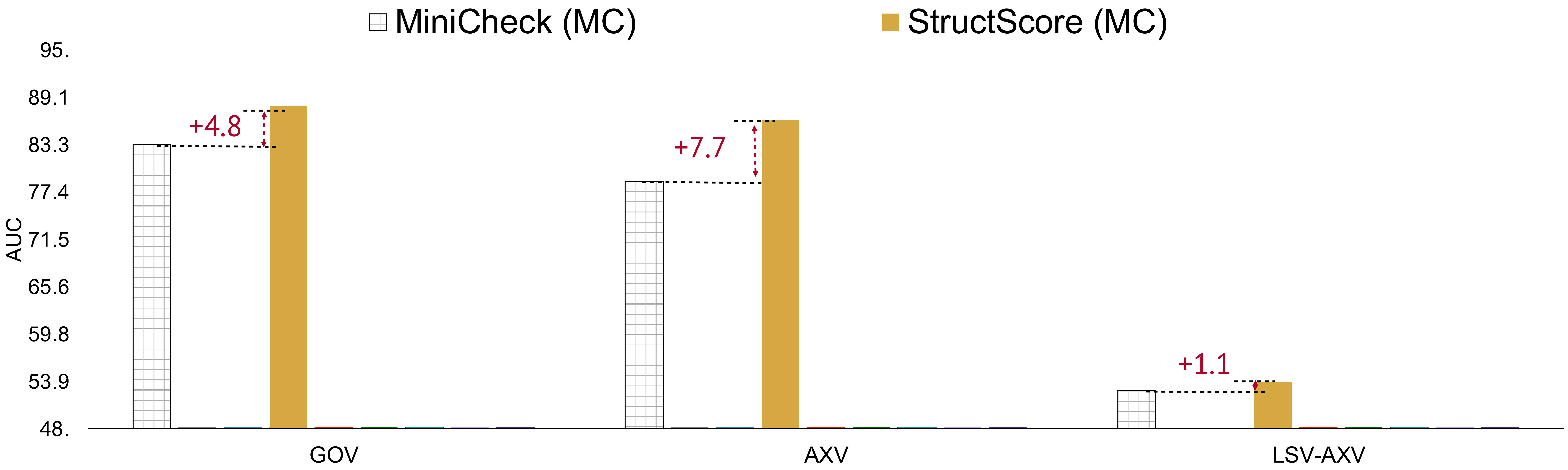
StructScore can outperform strong LLM-based baselines over several benchmarks.



# Results

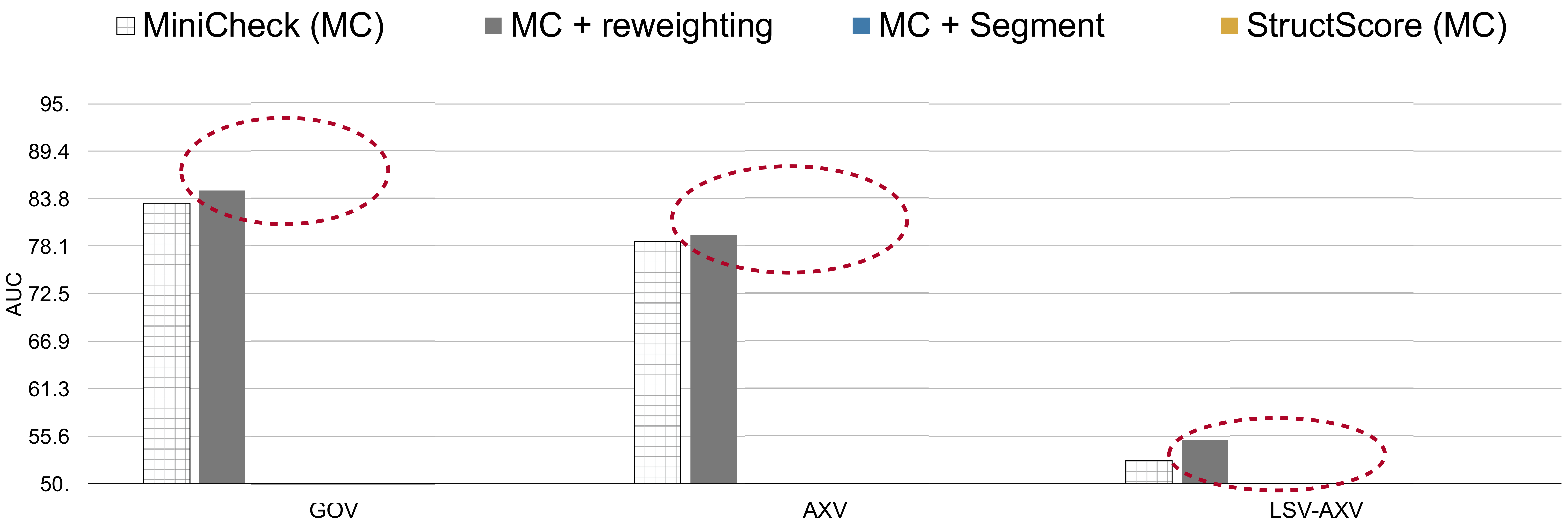
- StructScore enhances the backbone model by incorporating discourse-structure-inspired features.

to the



# Results

- StructScore enhances the backbone model by incorporating discourse-structure-inspired features
- Both source segmentation and the proposed reweighting algorithm can contribute to model performance, though their impact may vary.



# Takeaways

- Analysis of discourses level factors related to the factual inconsistency
  - *Finding 1: Sentences with complex structures are more prone to errors.*
  - *Finding 2: Errors are associated with the nuclearity and discourse features.*
  - *Finding 3: Discourse parsing facilitates long-doc segmentation by preserving structure.*
- The two components of StructScore enhance the backbone model at different levels.
- We hope our work can inspire continued exploration of discourse-level approaches for the evaluation of long document summarization.

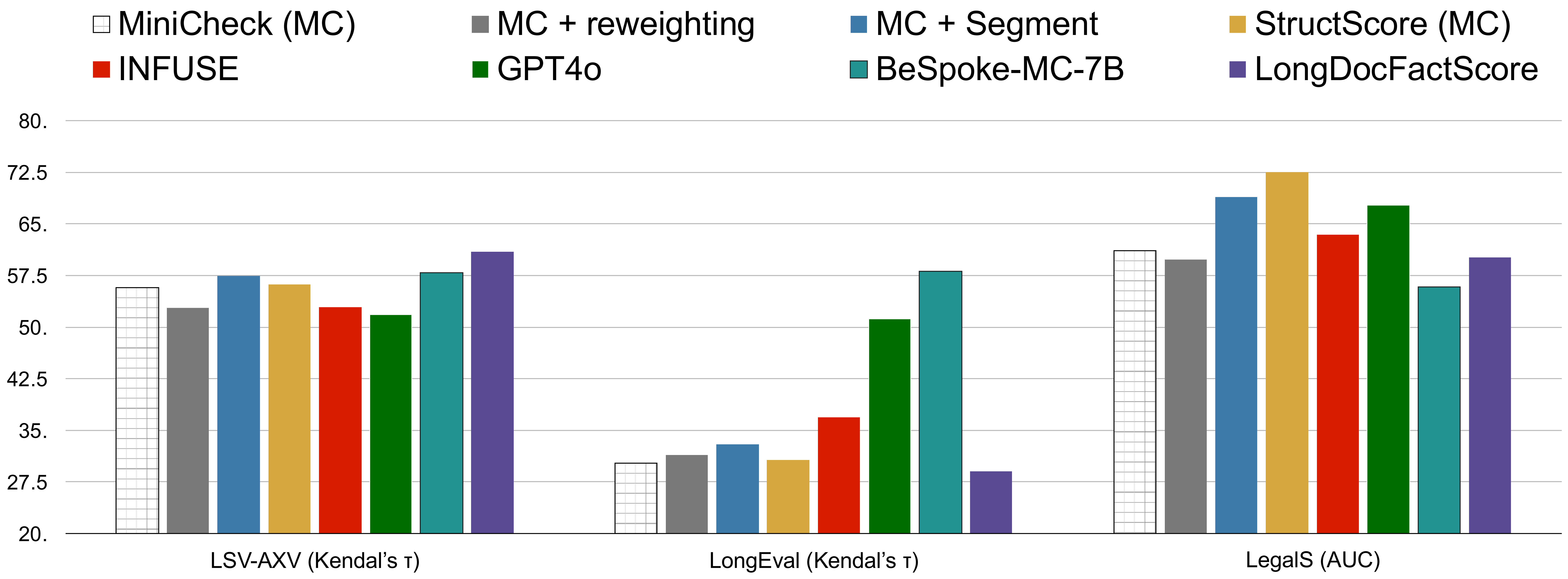
Thank you!

Discourse-Driven Evaluation:
Unveiling Factual Inconsistency
in Long Document Summarization

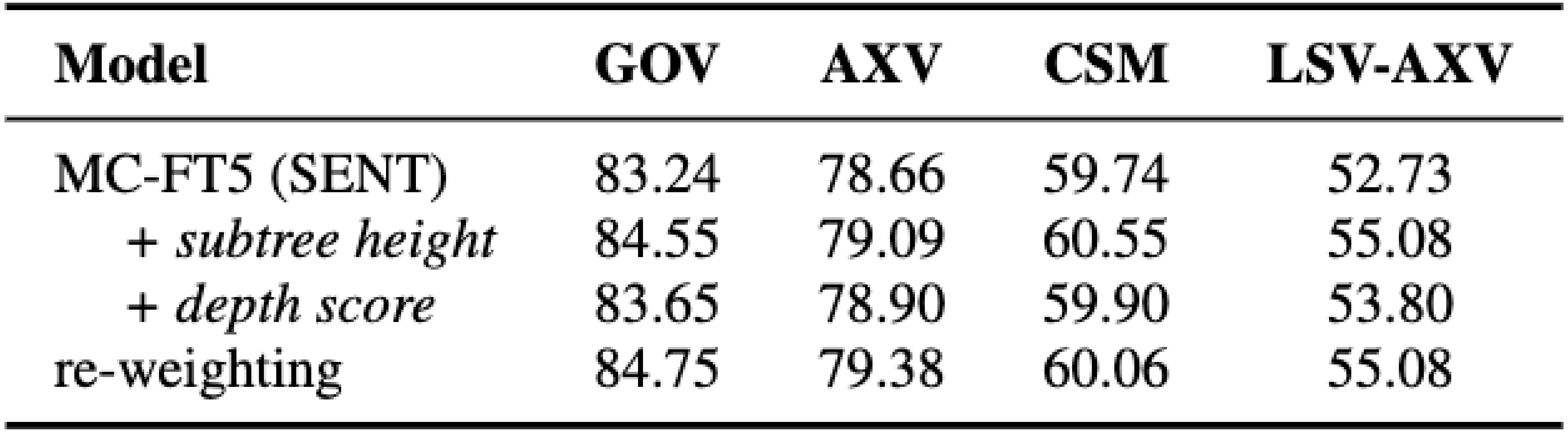
# Backup

# Extral Results

- There are also scenarios when the discourse-inspired approaches do not help.



# Ablation on Different Features



| **Model** | **GOV** | **AXV** | **CSM** | **LSV-AXV** |
|---|---|---|---|---|
| MC-FT5 (SENT) | 83.24 | 78.66 | 59.74 | 52.73 |
| + *subtree height* | 84.55 | 79.09 | 60.55 | 55.08 |
| + *depth score* | 83.65 | 78.90 | 59.90 | 53.80 |
| re-weighting | 84.75 | 79.38 | 60.06 | 55.08 |