Explanatory Summarization with Discourse-Driven Planning

Dongqi Liu[Ω], Xi Yu[Ω], Vera Demberg[Ω], Mirella Lapata[Θ]

[Ω]Saarland University, Germany

[Θ]University of Edinburgh, The United Kingdom

dongqi.me@gmail.com

Accepted by the Trans. of the Association for Computational Linguistics (TACL 2025)

European Research Council
Established by the European Commission

TL;DR

We propose EDU-level rhetorical planning using discourse structure and question-based cues to control explanatory content generation in lay summarization.

Motivation

- Why Do Lay Summaries Need Explanations?
- What Are Current Summarization Models Missing?
- Why Is Discourse-driven Planning a Promising Solution?
- What Are the Challenges We Address?

Motivation

- *Why Do Lay Summaries Need Explanations?*
  - Scientific concepts presented in academic documents are often too complex for non-experts to understand
  - Human-written lay summaries often contain analogies, causal justifications, and background
  - Just simplifying language (e.g., shorter sentences) may lead to loss of meaning or misinterpretation

Motivation

- *What Are Current Summarization Models Missing?*
  - Many models treat summarization as a flat end-to-end task without explicitly modeling explanations
  - Current models underproduce explanations, yielding summaries that are less clear, less accessible than human lay summaries

Motivation

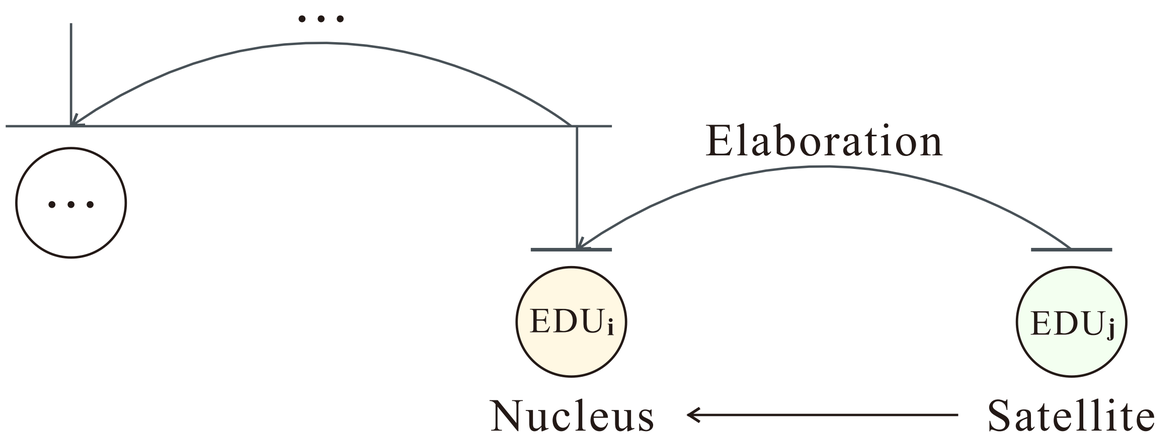
- *Why Is Discourse-driven Planning a Promising Solution?*
  - Discourse structures help to identify explanatory sentences and their rhetorical function
  - Planning offers controllability, enabling models to decide *what* and *where* to explain
  - Question-based plans naturally trigger explanation generation

Motivation

- *What Are the Challenges We Address?*
  - Lack of gold explanation annotations → need for automatic method (via discourse + LLMs)
  - Explanations are hard to evaluate automatically; many are misclassified as hallucinations

Prerequisite

- Rhetorical Structure Theory (RST)
  - Models discourse as a tree of Elementary Discourse Units (EDUs)
  - Connects via rhetorical relations
  - Defines nuclearity structure: nucleus (central) vs. satellite (supportive)
  - Reveals explanatory roles like *Justification* and *Background*

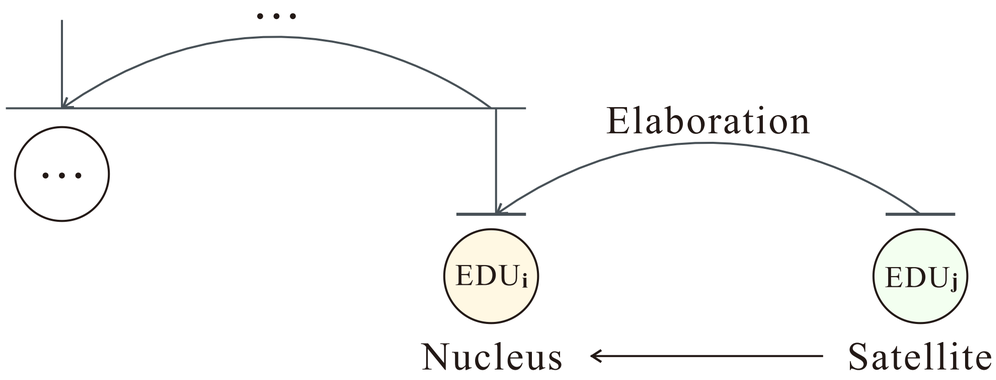


Prerequisite

- Question Under Discussion (QUD)
  - Models discourse via a stack of implicit questions
  - Each sentence resolves a current question in context
  - Adds an intentional layer to discourse modeling

Method Overview

- ***Objective:*** Generate lay summaries with explicit, controllable explanations
- ***Strategy:*** Use planning to guide both *where* and *how* to insert explanations
- ***Foundation:*** Leverage Rhetorical Structure Theory (RST) and the Question Under Discussion (QUD) framework



Lay Summary

... [The cerebellum utilizes proprioceptive feedback to fine-tune the timing of movements in a sequence based on previous actions.]^{EDU_i}[Imagine the cerebellum as a coach who watches how you perform a move, then gives tips to improve the next one based on what was seen.]^{EDU_j} ...

Source Document

... In principle, the cerebellar mechanisms underlying movement sequences could be different from, or at least somewhat different from those mediating single-component movements ... The cerebellum leverages proprioceptive feedback to calibrate the temporal dynamics of subsequent movements within a learned sequence. The cerebellum has been long implicated in learning and execution of accurate movements ... We employed electrical stimulation of mossy fibers as a cue to learn a single-component response.

Planning Pipeline

- ***Step 1:*** Automatically extract explanatory EDUs using DMRST parser
- ***Step 2:*** For each explanatory EDU, generate a corresponding plan question using GPT-4o
- ***Step 3:*** Construct a "**plan**" as an ordered list of questions, each prompting an explanation in the summary

Processed summary

[The cerebellum utilizes proprioceptive feedback to fine-tune the timing of movements in a sequence based on previous actions.]^{t_1} [Imagine the cerebellum as a coach who watches how you perform a move, then gives tips to improve the next one based on what was seen.]^{e_1} But how exactly does it achieve this? [To investigate, we trained rabbits to blink in response to an external cue and explored whether the cerebellum could use feedback from one blink to trigger the next.]^{t_2} [As expected, after learning the initial blink, the rabbits blinked again in response to their own first blink, creating a chain of movements.]^{e_2} Control experiments confirmed that each blink was initiated by the previous one rather than the original cue. Consistent patterns of brain activity during this process indicate that the cerebellum adjusts movement based on feedback from previous actions. [Building on this, we trained rabbits to blink on cue, and they learned to initiate additional blinks in response to earlier blinks in the sequence.]^{t_3} [We further found that the rabbits could use a blink from one eye as a cue to trigger a blink in the other eye, suggesting that the same mechanism governs these movements.]^{e_3} This raises the possibility that the cerebellum might also guide sequences of cortical activity during cognitive tasks, given its extensive connections to the cortex, a question future experiments should explore.

Target EDU Explanatory EDU

Planning questions

q1: How does the cerebellum use feedback to adjust the timing of movements in a sequence?
q2: How does the cerebellum use feedback from one blink to trigger the next in a sequence?
q3: How can a blink in one eye trigger a blink in the other eye?

Model Variants

- Plan-Output Model
  - Jointly generates plan questions and summary in a single sequence
- Plan-Input Model
  - Two-stage approach: first generate plan, then use it to guide summary generation

Experiments

- Three Lay Summarization Benchmarks

Dataset	# Training	# Validation	# Test	Avg. Doc Tokens	Avg. Summ Tokens	Coverage	Density	Compression Ratio
SciNews	33,497	4,187	4,188	7,760.90	694.80	0.74	0.94	12.71
eLife	4,346	241	241	7,833.14	383.02	0.82	1.77	20.52
PLOS	24,773	1,376	1,376	5,340.58	178.66	0.07	0.90	36.06

Experiments

- Discourse Parsing
  - RST Parser
    - Utilized DMRST for identifying explanatory EDUs and targets in reference summaries
    - (Full) Random replacement (FRR/RR) applied to simulate parser instability

Experiments

- **Alternative Parsing Strategies**
  - Rule-based extraction (Stede et al. 2017)
  - LLaMA-based RST parser (Maekawa et al., EACL 2024)
  - RST-Coref parser (Guz & Carenini, CODI 2020)
  - GPT-4o and Mistral as zero-shot extractors
- **Plan Generation**
  - GPT-4o generates plan questions for each explanatory EDU

Experiments

- Backbone
  - All candidate models built on Mistral-7B-Instruct-v0.3 (fully fine-tuned)
- Baselines
  - Backbone w/ zero-shot, in-context, and vanilla fine-tuning
- Other SOTAs

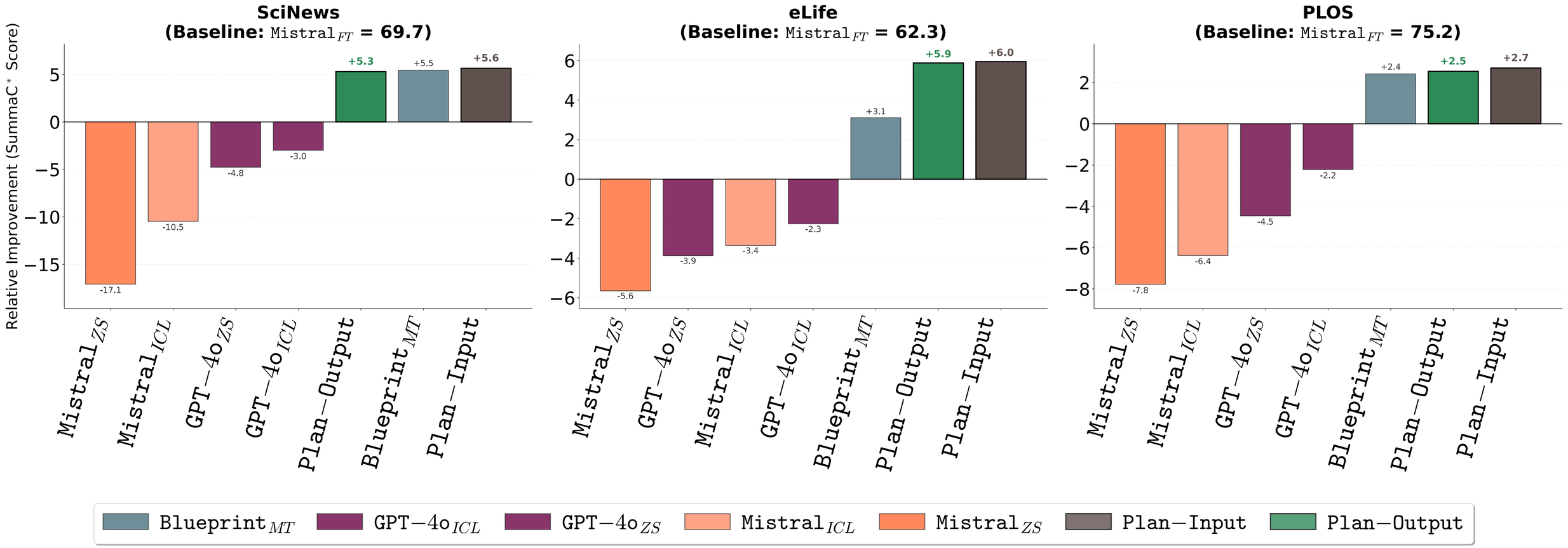
Experiments

- Automatic Metrics
  - **ROUGE-2**, **ROUGE-Lsum** (*informativeness*)
  - **BERTScore** (*semantic similarity*)
  - **D-SARI**, **FRE** (*readability*)
  - **ExpRatio** (*proportion of explanations*)
  - **SummaC**, **SummaC*** (*factual consistency*, incl. external verification)
  - **VeriScore** (*knowledge-grounded claim verification*)

Experiments

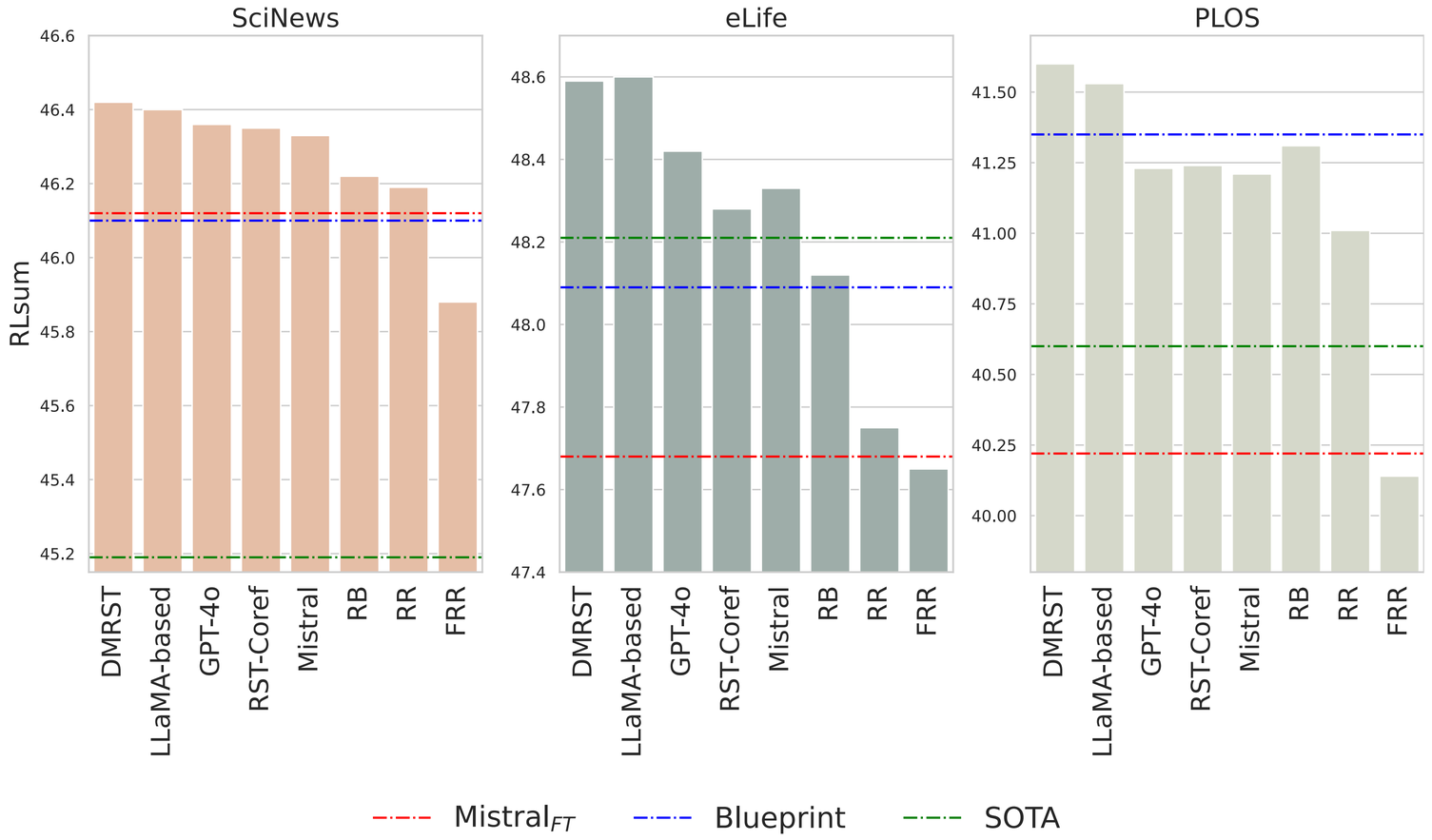
- Human & LLM-Based Evaluation
  - Six-dimension Likert scoring (*Faithfulness, Relevance, Informativeness, Accessibility, Explanation Accuracy, Explanation Usefulness*)
  - LLM-as-Judge (GPT-4o) for large-scale validation

Main Results



Parser Choice Analysis

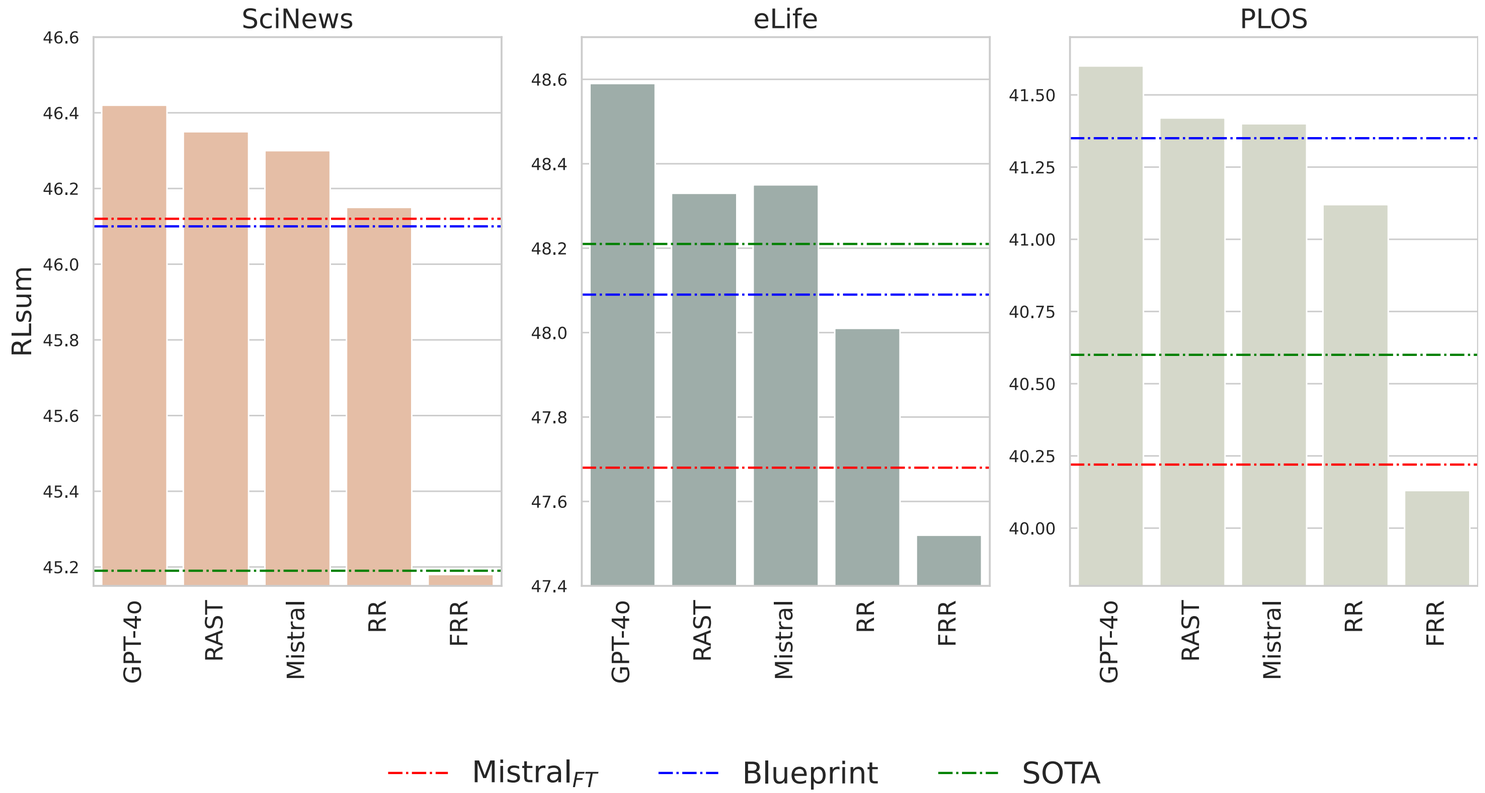
- Summary quality improves with more accurate discourse parsers
- DMRST parser gives best overall performance among other alternatives
- RST-based planning is robust to parser variation but degrades with random/noisy parsing



RR = random replacement, FRR = full random replacement

Plan Question Quality

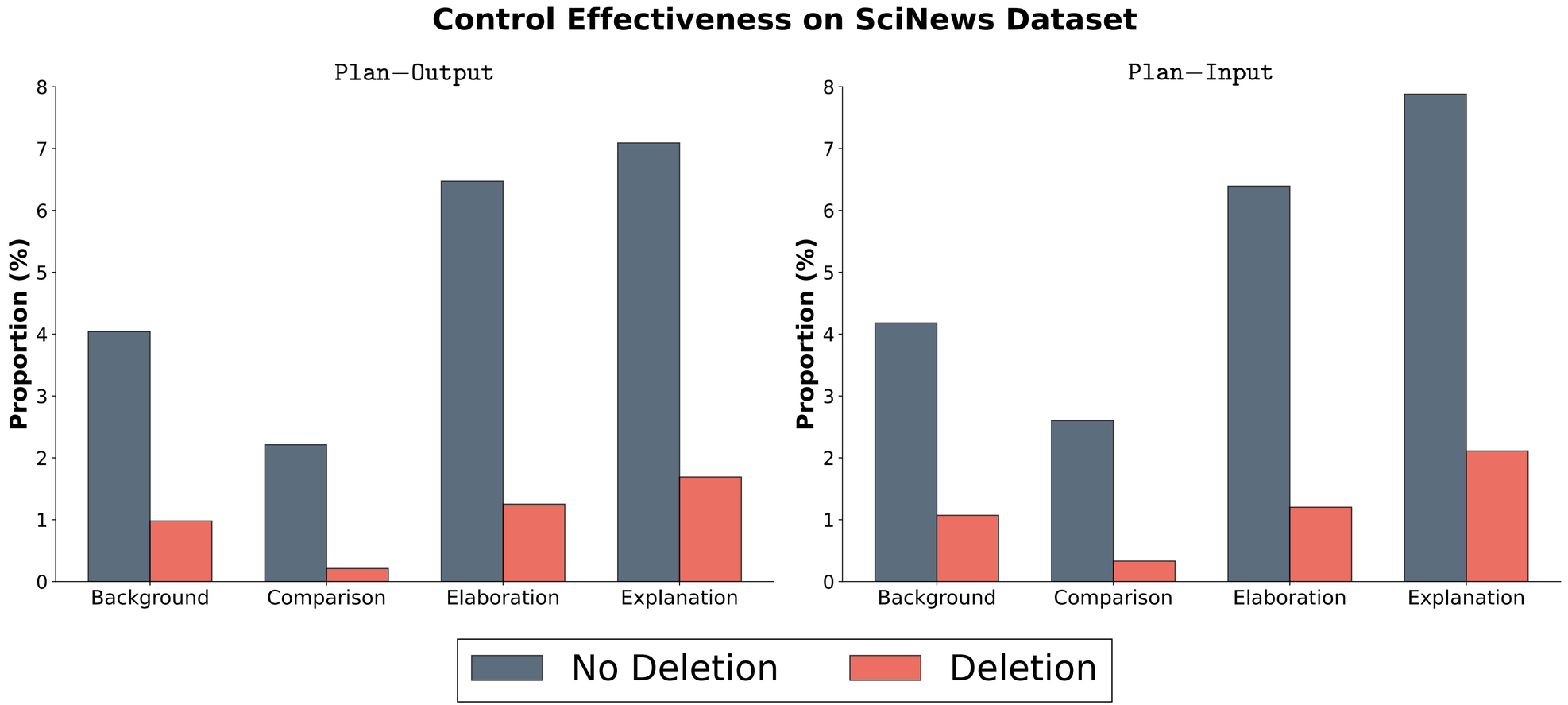
- Summary quality is directly impacted by the relevance of plan questions
- Robust to small noise, but random/irrelevant plans sharply reduce performance



RAST (Gou et al., EMNLP 2023) is a SOTA question generation method.
RR = random replacement, FRR = full random replacement

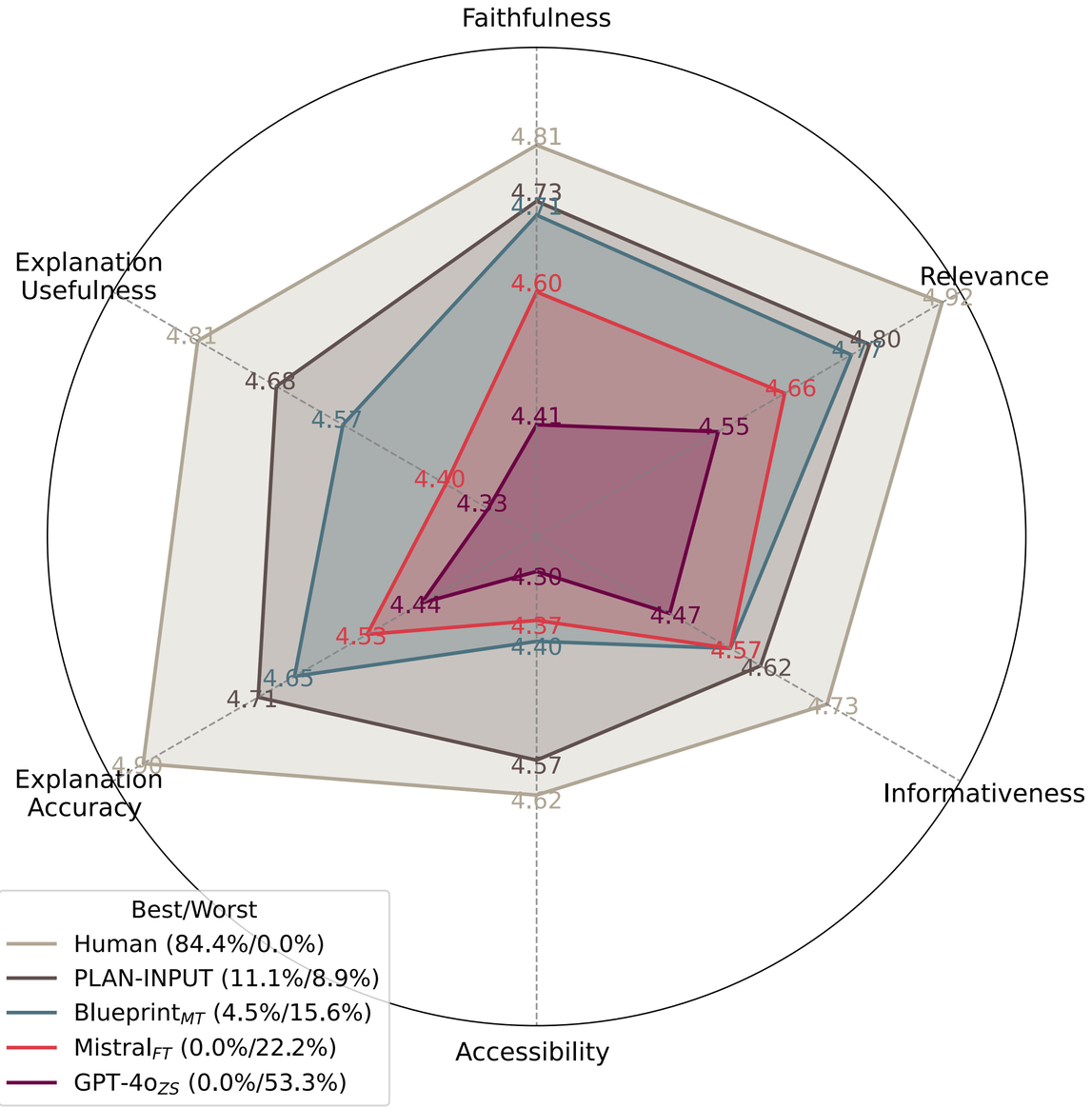
Controllability

- Removing plan questions for specific relations directly reduces corresponding explanation types in output



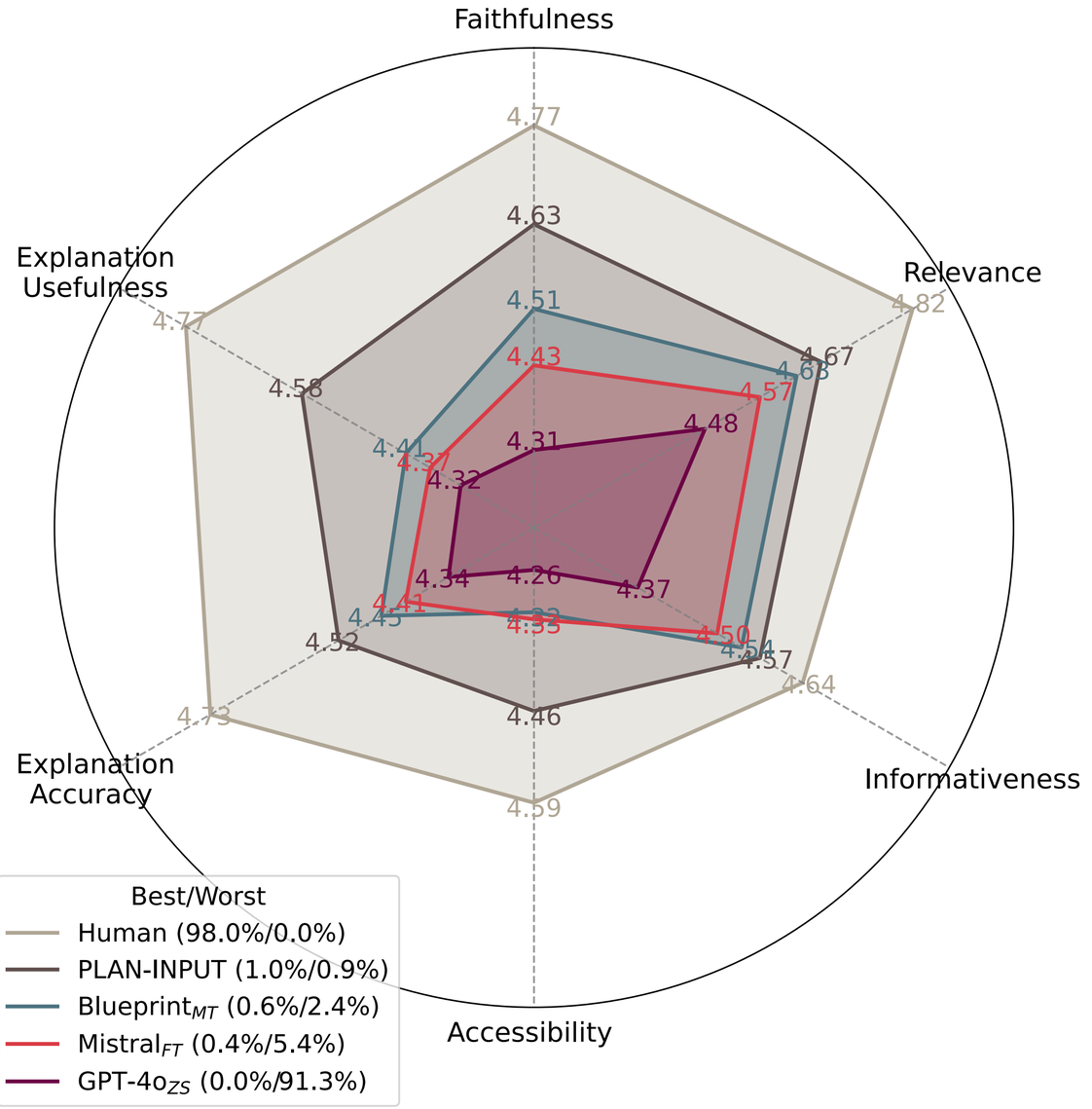
Human Evaluation

- Plan-Input achieves highest human ratings among neural models for faithfulness, relevance, informativeness, accessibility, and explanation usefulness
- Human-written summaries remain best overall, but Plan-Input is most competitive



LLM-as-Judge Evaluation

- LLM-based (GPT-4o) evaluation aligns with human ratings
- Plan-Input is consistently rated highest among models
- GPT-4o assigns lowest quality scores to its own generations



Conclusion

- We propose a discourse-driven, plan-based method that enables controlled generation of explanatory lay summaries.
- Our models achieve state-of-the-art performance in summary quality, factual consistency, and explanation diversity across multiple datasets.
- Planning at the EDU level allows fine-grained control and robust alignment with human-written summaries.

More Info

- **Code:** https://donggi.me/projects/ExpSum
- **Questions:** donggi.me@gmail.com

Thanks for listening

Q&A

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 948878). We acknowledge the inspiring environment of TRR 248 funded by DFG (German Research Foundation) – Project Number 389792660. Lapata acknowledges the support of the UK Engineering and Physical Sciences Research Council (Grant EP/W002876/1).