SciNews: From Scholarly Complexities to Public Narratives
A Dataset for Scientific News Report Generation

Dongqi Pu, Yifan Wang, Jia Loy, Vera Demberg

Department of Computer Science
Department of Language Science and Technology
Saarland Informatics Campus, Saarland University, Germany
dongqi.me@gmail.com

UNIVERSITÄT DES SAARLANDES

erc
European Research Council
Established by the European Commission

LREC-COLING 2024

TL;DR

- We introduce a **new corpus** designed to enhance the translation of **intricate research** into accessible scientific **news reports**

Motivation

- **Why** Study Scientific News Report Generation?
- **Similarities and Differences** with Summarization / Simplification

Motivation

- Why Study Scientific News Report Generation?
  - Academic publications → Require background knowledge 🤯
  - News reports → Increase accessibility with simplified language 😊

Academic Paper

Abstract Current techniques for characterizing cybersickness (visually induced motion sickness) in virtual environments rely on qualitative questionnaires. […]

Intro With the resurgence of virtual reality (VR), cybersickness has become […] We establish that cybersickness in an immersive HMD […] Our approach […] using inexpensive, commodity off -the-shelf devices for VR headsets and EEG devices. […] We find a **statistically significant correlation of Delta-, Theta-, and Alpha-waves with self-reported cybersickness.** […]

Conclusion Throughout the course of the study, we witnessed a wide range of reactions to the rendered stimuli. […] Our findings in this paper are just a first step to the many opportunities that present themselves in using EEG to study cybersickness in virtual environments. […] Finally, it will be highly **desirable, if at all possible, to move toward standards of assessing cybersickness and to use them to rate hardware (headsets, trackers, and displays) as well as the content (games, performances, and other immersive experiences).**

News Report

Report If a virtual world has ever left you **feeling nauseous or disorientated, you're familiar with cybersickness,** and you're hardly alone. The intensity of virtual reality (VR) whether that's standing on the edge of a waterfall in Yosemite or engaging in tank combat with your friends […] They were able to establish **a correlation between the recorded brain activity and self-reported symptoms** of their participants. […] **This helped the researchers identify which segments of the fly-through intensified users' symptoms.**

Motivation

- Similarities and Differences with Summarization / Simplification
  - Summarization: **Reduces** text, retains key content
  - Simplification: Uses **simpler** words/syntax for readability
  - Our task involves **both** simplifying and extracting

The SciNews Dataset

- Data Acquisition
- Data Cleaning
- Quality Control
    - Automated Quality Control
    - Human Quality Control
- Data Splits

The SciNews Dataset

- Data Acquisition
  - SciNews sourced from **Science X**
  - Selected open access articles with **CC-BY-4.0** license via DOI

https://sciencex.com/

The SciNews Dataset

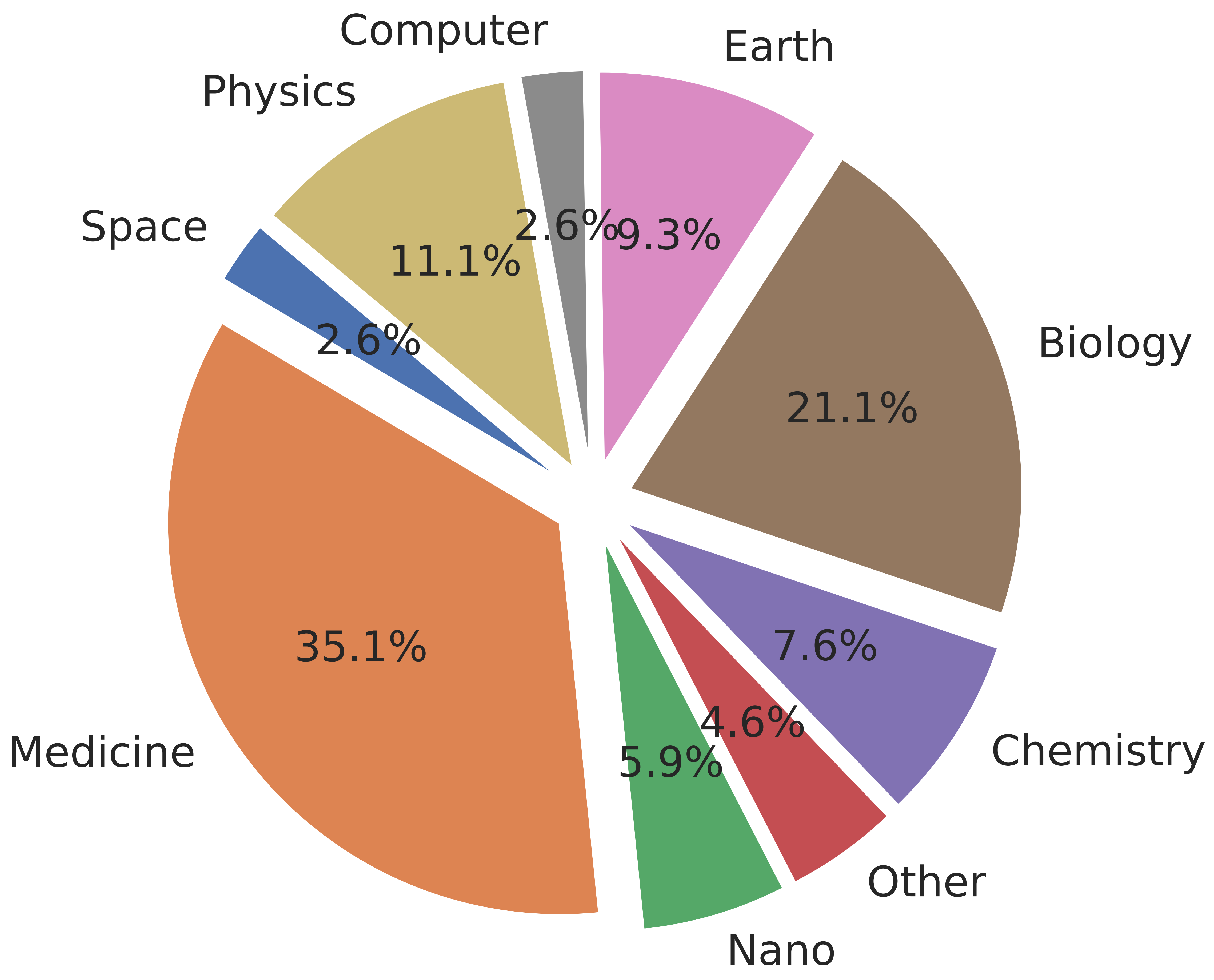
- Data Cleaning
  - Use PySBD and spaCy to clean texts; remove line breaks, emoticons, and links etc
  - Extract text from papers between the abstract and references
  - Exclude documents over 30,000 or under 2,000 words

The SciNews Dataset

- Quality Control
  - Automated Quality Control
    - Adapt methods from Mao et al. (2022) for vetting pairs; removed 612 of 42,484 pairs.
  - Human Quality Control
    - Inspired by Sun et al. (2021), we manually checked 100 sample pairs.

The SciNews Dataset

- Data Splits
  - 41,872 samples, split 80% training, 10% validation, 10% test across nine domains.



Dataset Analysis

- Dataset Comparison
- Dataset Statistics
- Papers vs. News

Dataset Analysis

- Dataset Comparison

  - SciNews vs. CSJ & PLOS: Similar sizes; SciNews has multidisciplinary labels.

  - Output Length: SciNews (695 tokens), PLOS (176 tokens), CSJ (361 tokens).

Dataset	Task	Language	Data Scope	Data Source	Scale	Input Level	Output Level	Multi-disciplinary?
LaySumm (Chandrasekaran et al., 2020c)	SLS	English	Archaeology, Hepatology, etc.	Research Papers	572	Document	Paragraph	✓
CDSR (Guo et al., 2021)	SLS	English	Healthcare	Research Papers	7805	Document	Paragraph	✗
CELLS (Guo et al., 2022)	SLS	English	Biomedicine	Research Papers	47157	Sentence	Sentence	✗
eLife (Goldsack et al., 2022)	SLS	English	Biomedicine	Research Papers	4828	Document	Paragraph	✗
PLOS (Goldsack et al., 2022)	SLS	English	Biomedicine	Research Papers	27525	Document	Paragraph	✗
SimpleScience (Kim et al., 2016)	STS	English	Biomedicine	Research Papers	293	Sentence	Vocabulary	✗
CLEAR (Grabar and Cardon, 2018)	STS	French	Biomedicine	Research Papers	663	Sentence	Sentence	✗
PLS (Devaraj et al., 2021)	STS	English	Medicine	Research Papers	4459	Paragraph	Paragraph	✗
SimpleText (Ermakova et al., 2022, 2023)	STS	English	Medicine & Computer Science	Research Papers	648	Sentence	Sentence	✓
CSJ (Fatima and Strube, 2023)	STS	English & German	Astronomy, Biology, etc.	Wikipedia	50132	Document	Paragraph	✓
SciNews (ours)	SNG	English	Science & Technology & Medicine	Research Papers	41872	**Document**	**Document**	✓

Dataset Analysis

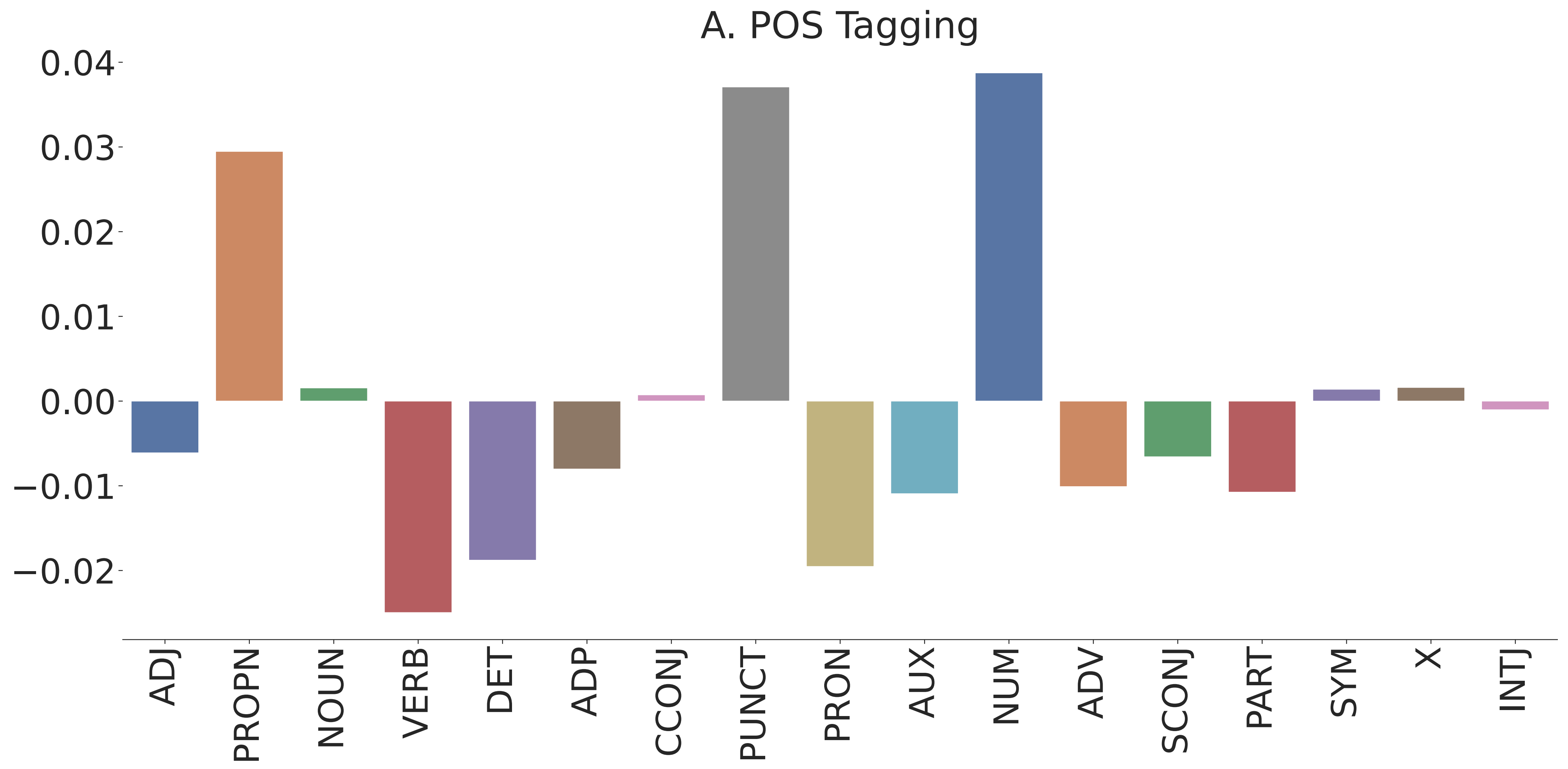
- Dataset Statistics
  - Long input & long output
  - High abstractive
  - High 1/2/3/4-grams novelty

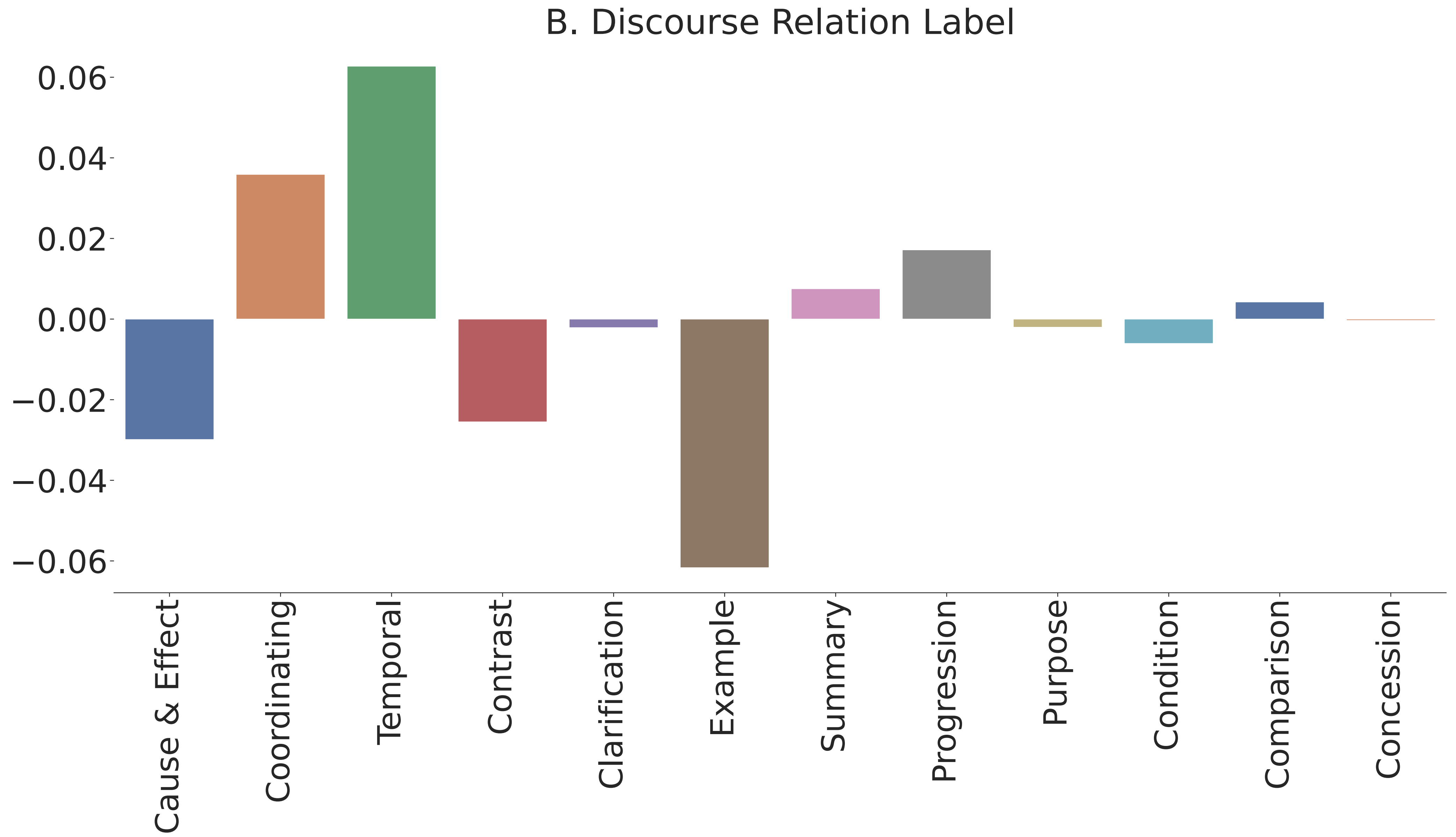
Property	Value
# Training Set	33497
# Validation Set	4187
# Test Set	4188
Avg. # Tokens (Papers)	7760.90
Avg. # Tokens (News)	694.80
Avg. # Sents. (Papers)	290.52
Avg. # Sents. (News)	25.17
Compression Ratio	12.71
Coverage	0.74
Density	0.94
1-gram Novelty	0.52
2-gram Novelty	0.91
3-gram Novelty	0.98
4-gram Novelty	0.99

Dataset Analysis

- Papers vs. News
  - First-person vs. third-person
  - **Lexical diversity:** Higher in news
  - **Syntax:** Simpler in news

Property	Papers	News
Type-Token Ratio↑	0.20	0.44
Lexical Density↑	0.42	0.46
Avg. # Difficult Words↓	773.08	134.84
Avg. # Modifiers per Noun Phrase↓	0.58	0.51
Avg. Depth of Dep Tree↓	6.94	6.25
FKGL↓	14.57	13.31
ARI↓	17.94	16.32





Experiments

- Baseline Models
- Experimental Settings
- Automatic Metrics

Experiments

- Baseline Models
  - Extractive Methods
    - Lead-3/K, Tail-3/K, and Random-3/K
    - Latent Semantic Analysis, LexRank, TextRank, Ext-oracle, and PacSum
  - Abstractive Methods
    - Longformer, RSTformer, SIMSUM (Seq2Seq)
    - Vicuna7B-16k, GPT-4 (GPT)

Experiments

- Experimental Settings
  - **Default Settings:** Use original model sizes, batch sizes, optimizers etc
  - **Decoding:** Beam search=3, trigram blocking, temperature=1, top-p=1
  - **Vicuna Model:** 5e-5 initial rate, cosine schedule, Adam optimizer, fine-tune 30 epochs

Experiments

- Automatic Metrics
  - F1 scores of Rouge-1 (R1), Rouge-2 (R2), Rouge-L (RL), and Rouge-Lsum (RLsum) (Lin, 2004)
  - BERTScore (Zhang et al., 2020)
  - METEOR (Banerjee and Lavie, 2005)
  - sacreBLEU (Post,2018)
  - NIST (Lin and Hovy, 2003)
  - SARI(Xu et al.,2016)

Results and Analysis

- General Results
- Comparison with Human-authored News Articles
- Automatic Inconsistency Detection
- Human Evaluation
- GPT-4 Evaluation
- Model Errors

Results and Analysis

- General Results

Model	$R1_{f1}$↑	$R2_{f1}$↑	RL_{f1}↑	$RLsum_{f1}$↑	$BERTscore_{f1}$↑	Meteor↑	sacreBLEU↑	NIST↑	SARI↑
Full article (lower bound)	14.42	5.21	6.90	13.94	58.55	0.21	1.49	0.55	34.83
Lead-3	14.65	4.47	8.93	13.47	54.69	0.06	0.12	0.00	35.79
Lead-K	41.99	10.96	16.13	39.68	58.55	0.27	5.25	2.34	37.21
Tail-3	8.43	1.46	5.41	7.77	43.61	0.03	0.05	0.01	33.94
Tail-K	32.16	5.58	13.37	30.49	51.83	0.20	2.16	1.76	35.50
Random-3	10.20	1.84	6.43	9.30	47.68	0.04	0.05	0.01	34.23
Random-K	35.91	6.90	14.10	33.83	54.41	0.22	2.68	1.97	35.94
LSA (Steinberger et al., 2004)	39.75	8.45	15.10	37.40	56.43	0.25	3.42	2.19	36.13
LexRank (Erkan and Radev, 2004)	35.59	7.98	14.97	33.62	54.49	0.24	3.22	1.92	36.16
TextRank (Mihalcea and Tarau, 2004)	35.64	7.85	14.77	33.52	53.80	0.23	3.17	1.94	36.13
PacSum (Zheng and Lapata, 2019)	41.03	10.53	15.47	38.75	57.64	0.27	4.82	2.28	36.85
Ext-oracle (Narayan et al., 2018)	42.58	11.92	16.16	40.38	56.60	**0.30**	5.90	2.43	37.28
$GPT\text{-}4_{ZS}$ (OpenAI, 2023)	41.38	9.03	15.25	39.01	58.33	0.19	4.64	1.12	37.52
SIMSUM (Blinova et al., 2023)	44.38	12.20	18.13	41.46	60.09	0.27	6.31	2.38	40.54
Longformer (Beltagy et al., 2020)	47.60	14.74	19.09	44.83	62.84	0.28	7.64	2.47	41.52
RSTformer (Pu et al., 2023)	**48.21**‡	**14.92**	**20.12**‡	**45.19**‡	62.80	0.28	**7.70**	**2.55**	41.56
Vicuna7B-16k (Zheng et al., 2023)	47.75	14.88	19.92	45.01	**62.88**	**0.30**	7.69	2.53	**41.71**‡

Table 4: Model performance. The bold numbers represent the best results with respect to the given test set. ‡ denotes that the value is significantly superior to those of all other models according to the Wilcoxon signed-rank test in the corresponding indicator ($p<0.05$).

Results and Analysis

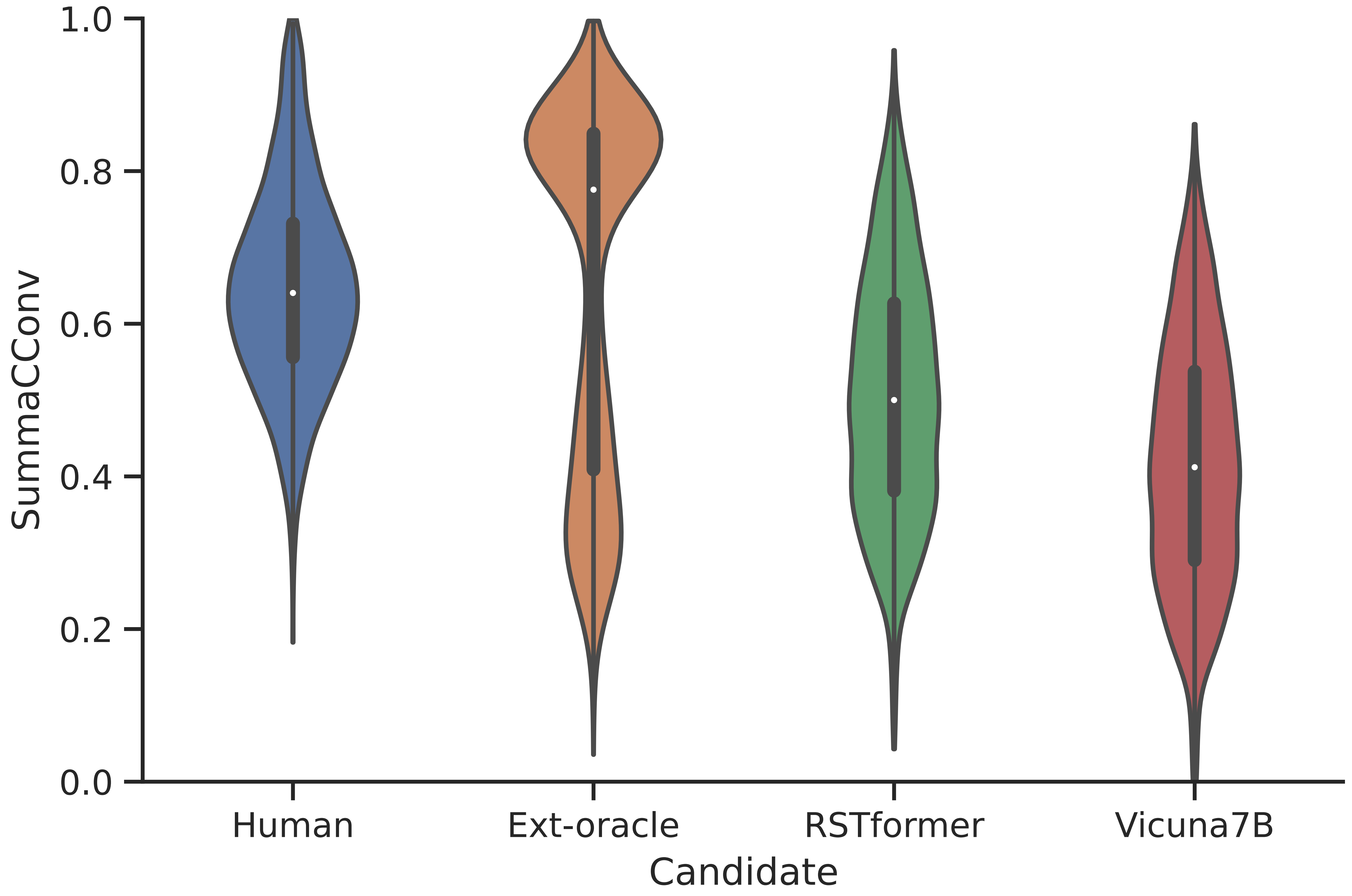
- Comparison with Human-authored News Articles
  - **Lexical Diversity:** RSTformer closest to human.
  - **Complexity:** Vicuna generates longer, complex words.
  - **Readability:** Humans outperform models (FKGL, ARI).

Metric	Human	Ext-oracle	RSTformer	Vicuna7B
Avg. # Tokens	696.19	1274.54	653.37	782.21
Avg. # Sents.	25.29	44.51	22.85	25.03
Type-Token Ratio↑	0.45	0.40	**0.47**	0.37
Lexical Density↑	**0.46**	0.44	**0.46**	0.42
Avg. # Difficult Words↓	**134.65**‡	217.37	141.75	164.5
Avg. # Modifiers per NP↓	**0.50**	0.61	0.57	0.62
Avg. Depth of Dep Tree↓	**6.24**‡	6.68	7.62	6.72
FKGL↓	**13.27**‡	15.80	14.95	14.12
ARI↓	**16.26**‡	19.20	18.22	16.90

Table 5: Models vs. Humans; ‡ indicates that the value significantly differs from those of all other candidates in the same test set, according to the Wilcoxon signed-rank test for the corresponding indicator ($p<0.05$).

Results and Analysis

- Automatic Inconsistency Detection
  - Abstractive models lower than humans; extractive highest.



Results and Analysis

- Human Evaluation
  - **Evaluation Setup:** 10 samples, 4 candidate reports, blind testing by Masters/PhD evaluators.
  - **Criteria:** Relevance, simplicity, conciseness, faithfulness; scored 1-3.
  - **Results:** RSTformer and Vicuna excel in different areas; overall, models lag behind human proficiency.

Candidate	Relevant	Simple	Concise	Faithful	Best | Worst
Human	**2.67**/$_{0.23}$	**2.83**‡/$_{0.33}$	**2.43**‡/$_{0.33}$	**2.73**‡/$_{0.10}$	70.00% | 3.33%
Ext-oracle	2.63/$_{0.33}$	1.30/$_{1.00}$	1.20/$_{1.00}$	2.63/$_{0.17}$	0.00% | 80.00%
RSTformer	2.63/$_{0.40}$	2.27/$_{0.67}$	2.03/$_{0.73}$	2.17/$_{1.00}$	20.00% | 3.33%
Vicuna7B	2.47/$_{0.60}$	2.47/$_{0.67}$	2.17/$_{0.60}$	1.96/$_{1.00}$	10.00% | 13.33%

Table 6: Human evaluation results: average ratings (on a scale from 1 to 3). The number following the slash represents the percentage of evaluation samples in which an issue identified by evaluators occurs at least once.

Results and Analysis

- GPT-4 Evaluation
  - Uses human evaluation guidelines, resets history for unbiased assessment.
  - **Preliminary Check:** GPT-4 and human scores align across criteria.
  - Overall Findings: Humans outperform models; extractive method often rated worst.

Candidate	Relevant	Simple	Concise	Faithful	Best | Worst
Human	**2.86**‡	**2.77**‡	**2.83**‡	**2.91**‡	92.00% | 0.00%
Ext-oracle	2.73	1.73	1.55	2.70	0.00% | 93.00%
RSTformer	2.69	2.41	2.42	2.47	6.00% | 2.00%
Vicuna7B	2.56	2.59	2.53	2.32	2.00% | 5.00%

Table 7: GPT-4 evaluation results on 100 samples

Results and Analysis

- Model Errors
  - Hallucinations
  - Factual Errors
  - Generalization

Conclusion

- **Dataset Introduction:** "SciNews" comprises 40,000+ scientific papers with paired news reports.
- **Exploratory Analysis:** Reveals challenges and research prospects for state-of-the-art models.
- **Dataset Potential:** Enhances scientific news generation, offers resource for NLP tasks like topic classification.

More Info

- **Data & Code:** https://dongqi.me/projects/SciNews
- **Questions:** dongqi.me@gmail.com

Thanks for listening

Q&A