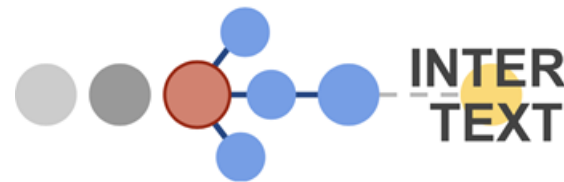




How to InterText? Modeling Text as a Living Object in Context



IG
2min ago

Dagstuhl Seminar
"Reviewer No. 2"

29.01.-02.02.24



A lot of work is text work



Humans are not good at handling lots of text



Yann LeCun @ylecun · Feb 19

Crypto bro: AI labs should stop publishing useless papers and make products that increase shareholder value.

Me: There would be no new product without the research published in those papers. New products aren't pulled out of thin air, unlike cryptocurrency value.

2,542 likes | 412.9K retweets

Abstract

Pre-trained text encoders have rapidly advanced the state of the art on many NLP tasks. We focus on one such model, BERT, and aim to quantify where linguistic information is captured within the network. We find

that the model represents traditional NLP pipeline localizable way, and is suitable for each step application: POS tagging, roles, then coreference reveals that the model just this pipeline dynamic level decisions on the information from high

7:28 16 Feb

Park funding to boost UK space sector

By Sonia Kataria
BBC News



The facility is to receive £284,000 from the UK Space Agency to fund a cluster development manager.

[Read more >](#)

Advances in Neural Information Processing Systems 34 (NeurIPS 2021)

Edited by: M. Ranzato and A. Beygelzimer and Y. Dauphin and P.S. Liang and J. Wortman Vaughan
ISBN: 9781713845393 [Purchase Printed Proceeding](#)

- Beyond Value-Function Gaps: Improved Instance-Dependent Regret Bounds for Episodic Reinforcement Learning *Christoph Dann, Teodor Vanislavov Marinov, Mehryar Mohri, Julian Zimmert*
- Learning One Representation to Optimize All Rewards *Ahmed Touati, Yann Olivier*
- Matrix factorisation and the interpretation of geodesic distance *Nick Whiteley, Annie Gray, Patrick Rubin-Delanchy*
- UniDoc: Unified Pretraining Framework for Document Understanding *Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalos, Ani Nenkova, Tong Sun*
- Finding Discriminative Filters for Specific Degradations in Blind Super-Resolution *Liangbin Xie, Xintao Wang, Chao Dong, Zhongang Qi, Ying Shan*
- Counterfactual Explanations Can Be Manipulated *Dylan Slack, Anna Hilgard, Himabindu Lakkaraju, Sameer Singh*
- From Canonical Correlation Analysis to Self-supervised Graph Neural Networks *Hengrui Zhang, Qitian Wu, Junchi Yan, David Wipf, Philip S Yu*
- BAST: Bayesian Additive Regression Spanning Trees for Complex *Sang, Bani Mallick*
- Hyperbolic Busemann Learning with Ideal Prototypes *Mina Ghadimi*
- Backward-Compatible Prediction Updates: A Probabilistic Approach *Matthäus Kleindessner, Francesco Locatello, Bernhard Schölkopf*
- Truncated Marginal Neural Ratio Estimation *Benjamin K Miller, Akshay Krishnamoorti, Christoph Weniger*
- ReAct: Out-of-distribution Detection With Rectified Activations *Y*



WIKIPEDIA
The Free Encyclopedia

English 6 606 000+ articles	Русский 1 887 000+ статей
日本語 1 359 000+ 記事	Deutsch 2 764 000+ Artikel
Français 2 488 000+ articles	Español 1 833 000+ artículos
Italiano 1 792 000+ voci	中文 1 331 000+ 条目 / 条目
فارسی 947 000+ مقاله	Polski 1 552 000+ hasel




Some things got much easier...



TECHNISCHE
UNIVERSITÄT
DARMSTADT

EN  Search

Translation

English ↔ Georgian

I would like a coffee please

ყავა მინდა
q'ava minda

How to I bake a cake without flour, eggs, sugar and water

and water, as these ingredients are commonly used in cakes to provide structure, leavening, and sweetness. However, there are some alternative ingredients that you could use to try to make a cake. For example, you could use a combination of ground nuts or oats as a flour alternative, and mashed bananas or applesauce as a binding agent in place of eggs. To add sweetness to the cake, you could try using a natural sweetener like honey or maple syrup, or you could try using fruit purees or unsweetened applesauce. Instead of using water, you could try using a non-dairy milk like almond milk or coconut milk. Keep in mind that using these alternative ingredients may result in a cake that is different in texture and flavor than a traditional cake.

Information extraction

BBC News 

Broadcasting corporation

bbc.co

BBC News is British Broadcasting Corporation responsible for the gathering and broadcasting of news and current affairs in the UK and around the world. [Wikipedia](#)

Headquarters: London, United Kingdom

Parent organization: BBC Public Service Broadcaster

Founded: November 14, 1922, Central London, London, United Kingdom

Key people: Deborah Turness (Director of News & Current Affairs), Mary Hockaday (Head of Newsroom)

Number of employees: 3,500 (2,000 journalists)

Recommendations

New Papers From Your Alerts Manage Alerts

New Papers From Luke Zettlemoyer

Representation Deficiency in Masked Language Modeling
Yu Meng, Jitin Krishnan, +6 authors [Luke Zettlemoyer](#) · Computer Science · ArXiv · 4 February 2023
Masked Language Modeling (MLM) has been one of the most prominent approaches for pretraining bidirectional text encoders due to its simplicity and effectiveness. One notable concern about MLM is that... [Expand](#)

Toolformer: Language Models Can Teach Themselves to Use Tools
Timo Schick, Jane Dwivedi-Yu, +5 authors [Thomas Scialom](#) · Computer Science · ArXiv · 9 February 2023
Language models (LMs) exhibit remarkable abilities to solve new tasks from just a few examples or textual instructions, especially at scale. They also, paradoxically, struggle with basic... [Expand](#)

Transformers and LLMs

...but some things are still hard



Fake news and
Information tracing

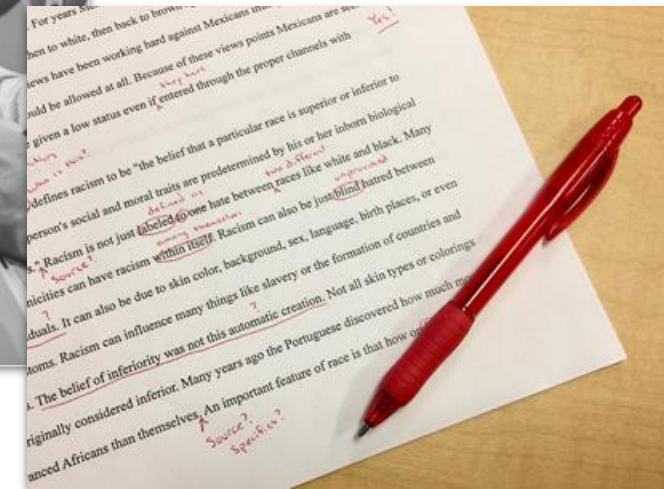


source?

Quality control



Collaborative writing



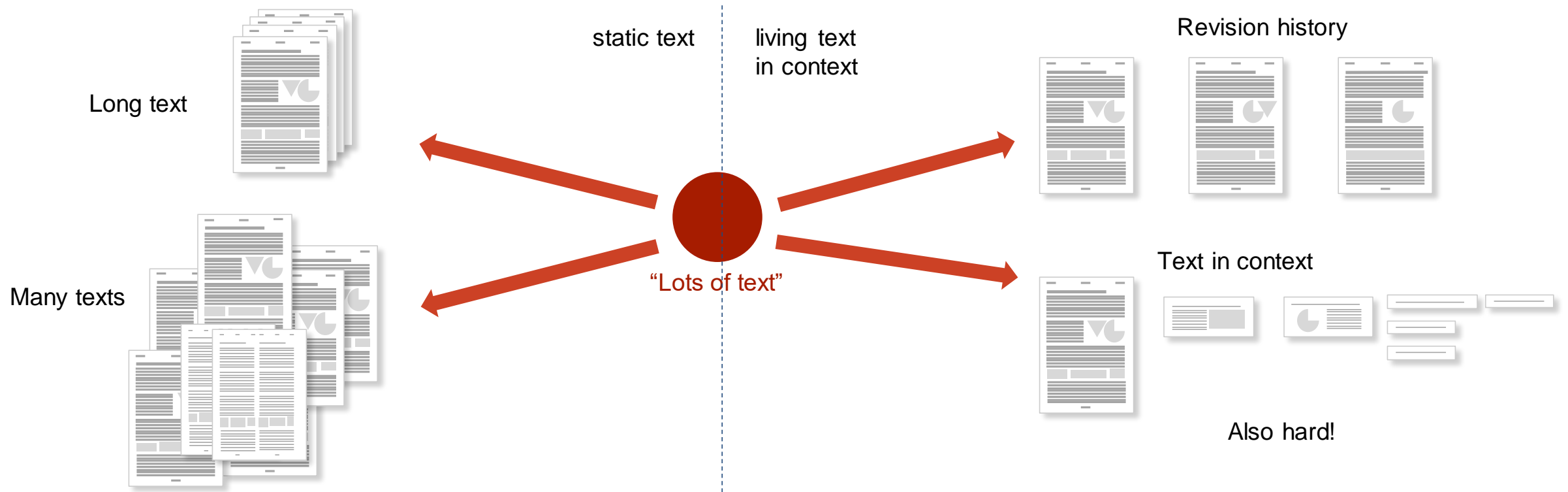
Abstract

Pre-trained text encoders have rapidly advanced the state of the art on many NLP tasks. We focus on one such model, BERT, and aim to quantify where linguistic information is captured within the network. We find that the model represents the steps of the traditional NLP pipeline in an interpretable and localizable way, and that the regions responsible for each step appear in the expected sequence: POS tagging, parsing, NER, semantic roles, then coreference. Qualitative analysis reveals that the model can and often does adjust this pipeline dynamically, revising lower-level decisions on the basis of disambiguating information from higher-level representations.

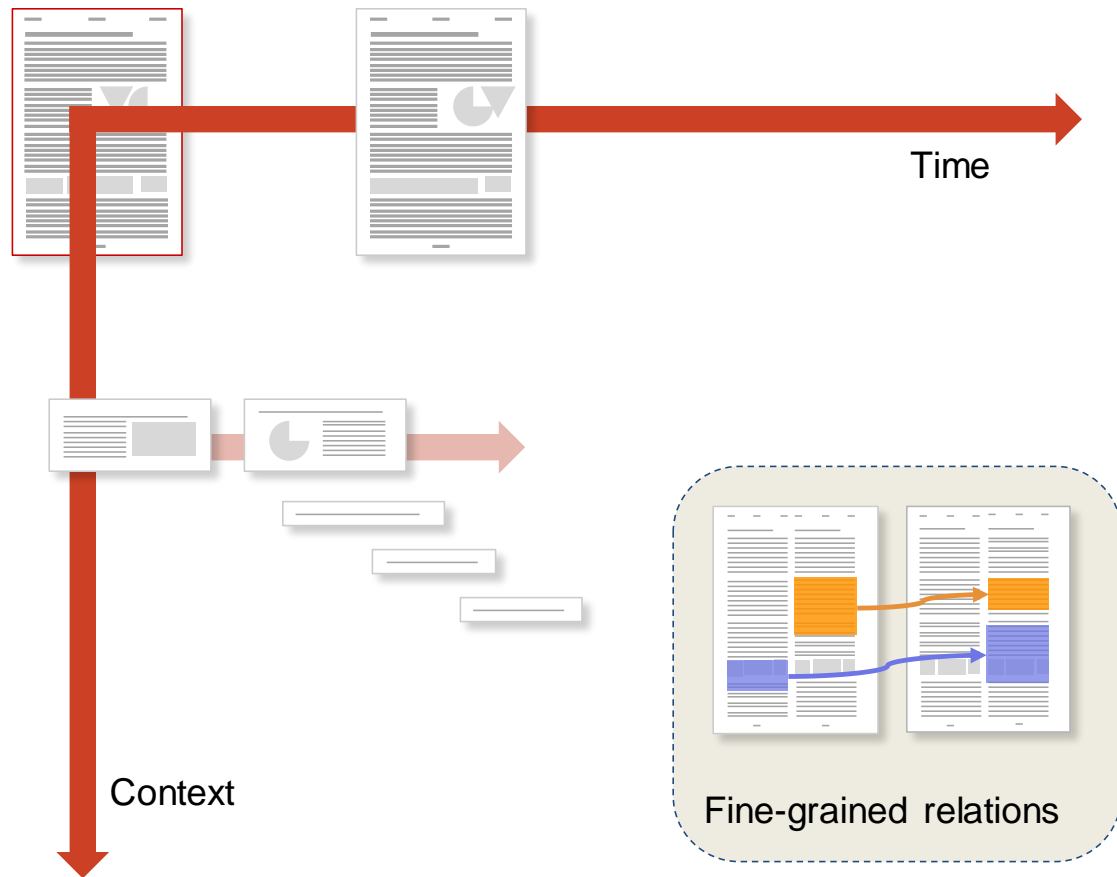
Assisted reading

This was discussed in a recent paper: <...>

Why?

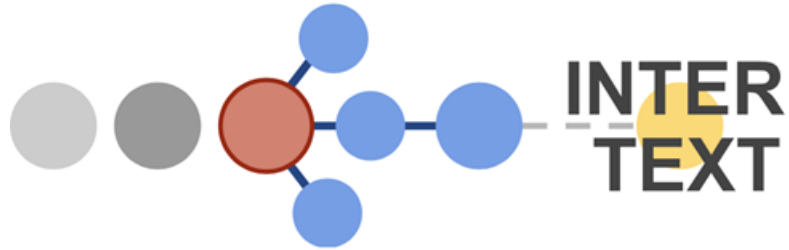


The InterText initiative



- Text as a living object in context
- Inspired by intertextuality theories
- Conceptual models
- Tasks
- Methods
- Datasets
- Across domains and text genres

The InterText initiative



- Several coordinated research projects
- Five-year ERC AdG



Iryna Gurevych
Principal Investigator



Ilia Kuznetsov
Postdoc



Martin Tutek
Postdoc



Jan Buchmann
PhD Student



Nils Dycke
PhD Student



Max Eichler
PhD Student



Dennis Zyska
PhD Student



Qian Ruan
PhD Student

... and many
more!

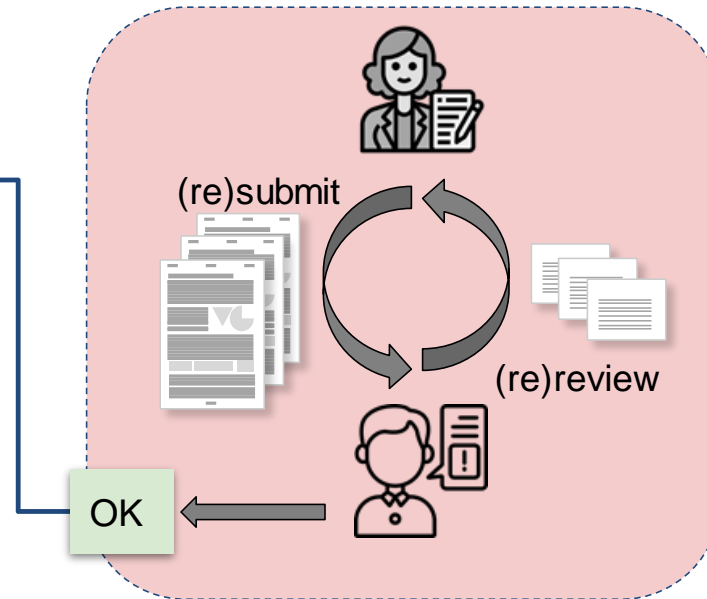


<https://intertext.ukp-lab.de/>

Case Study: Peer Review

Scientific documents

- lots of text
- reference
- versioning
- many great applications

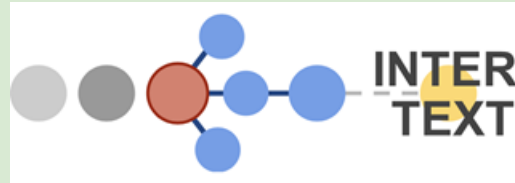


Peer review

- turn-based text discussion
- turn-based text editing
- closed environment
- needs help

→ **New tasks, data and methods
in NLP for peer review**

Big picture



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Long-document processing

How to incorporate (cross-)document structure into language modeling?

- *Document Structure in Long Document Transformers* EACL-2024
- *HDT: Hierarchical Document Transformer* (+ Uni Tübingen) COLM-2024, to appear

Reading and writing assistance

How do people read and write texts, and how can AI help them do it better?

- *CARE: Collaborative AI-Assisted Reading Environment* ACL 2023
- 2 x ACL 2024
- [Ongoing]



Intertextual modeling

What relations can hold between texts, and how to model them?

This talk

AI to support peer review

How can AI make scholarly peer review more efficient?

- *What Can Natural Language Processing Do for Peer Review?* arXiv, 2024
- *Jiu-Jitsu Argumentation for Writing Peer Review Rebuttals.* EMNLP-2023
- *Does My Rebuttal Matter? Insights from a Major NLP Conference.* NAACL-2019

<https://intertext.ukp-lab.de/>

Talk overview



NLPeer

F1000RD

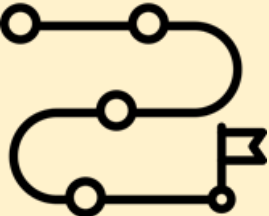
Linking / Versioning

Data & Tasks

Related work generation

Rebuttal generation

Text Generation



Discussion and Outlook

Intertextual modeling
What relations can hold between texts, and how to model them?

This talk



Data & Tasks

Text Generation

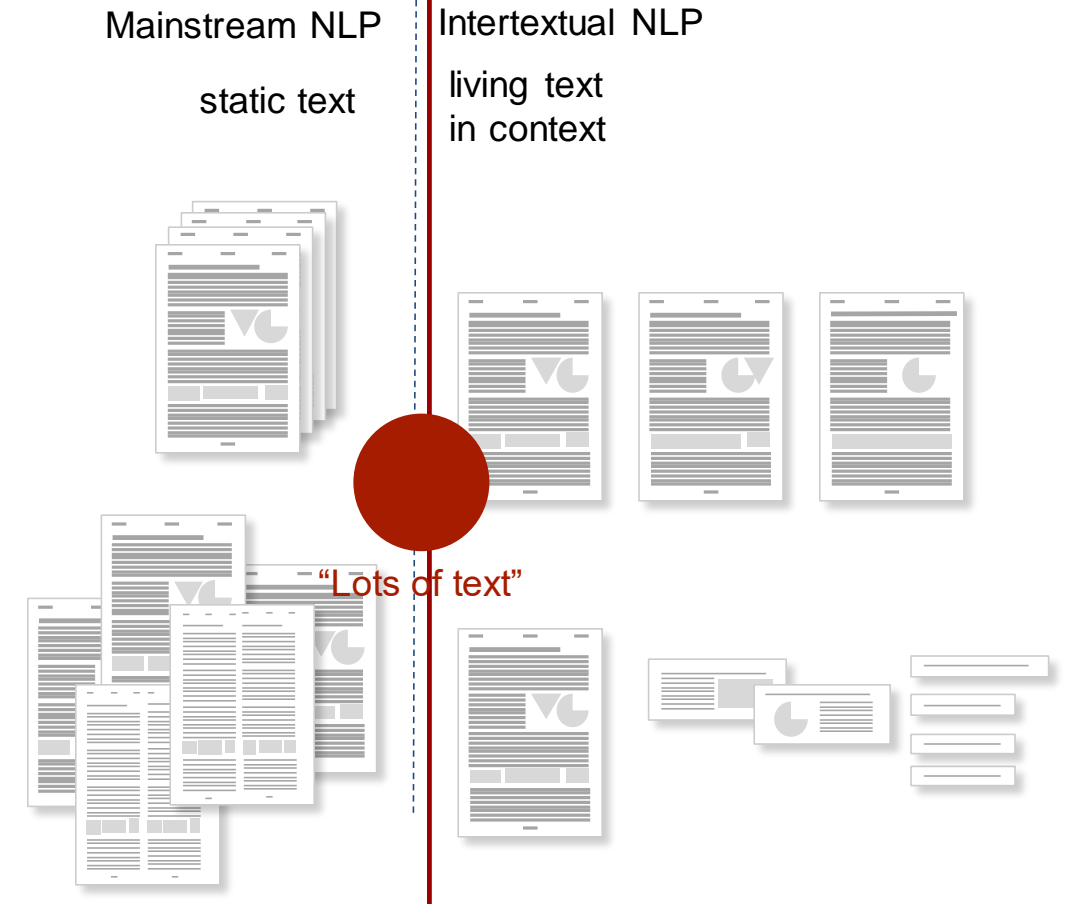
Outlook

Q&A

Data and Tasks

Getting the data

- Ethics, licensing, GDPR
- Static texts are easier
 - One source
 - One (group of) authors
 - Established genres with clear rules
- Living texts in context are grey zone
 - New document types
 - Attribution and personal data
 - Confidentiality



Getting the data: Workflow

- Three approaches

- Just take the data

1

- Pros: easy
- Cons: bad

- Look for open sources

2

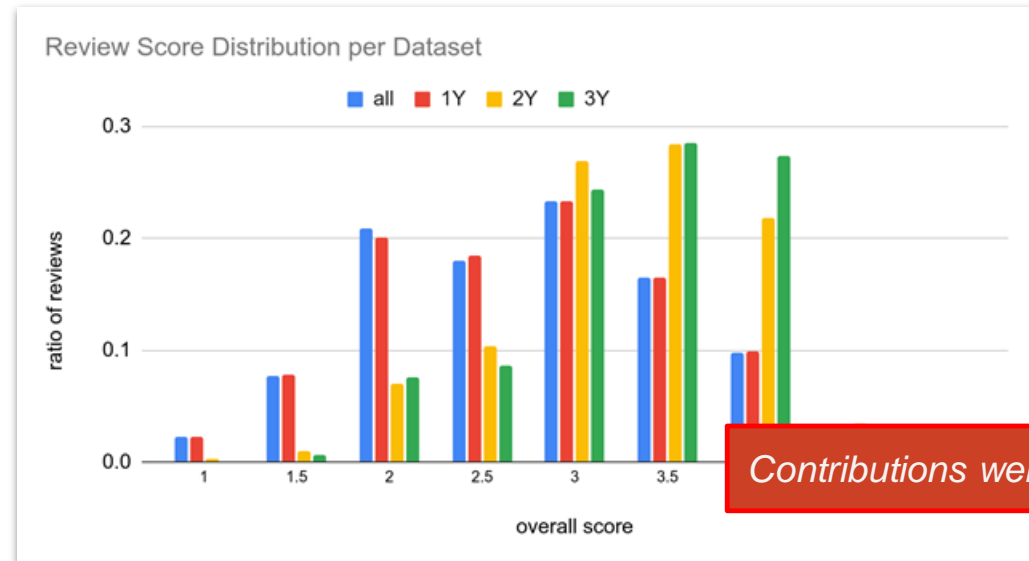
- Pros: easy
- Cons: limiting

- Data donation

3

- Pros: any data source
- Cons: needs some work

The “Yes-Yes-Yes” Donation Workflow at ACL ARR



Dycke et al. 2022 “Yes-Yes-Yes: Proactive Data Collection for ACL Rolling Review and Beyond”. Findings of EMNLP-2022

Getting the data: F1000RD



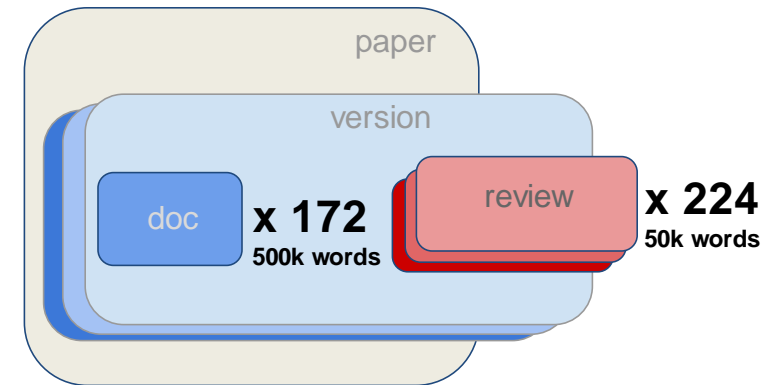
F1000RD

First corpus for intertextual NLP

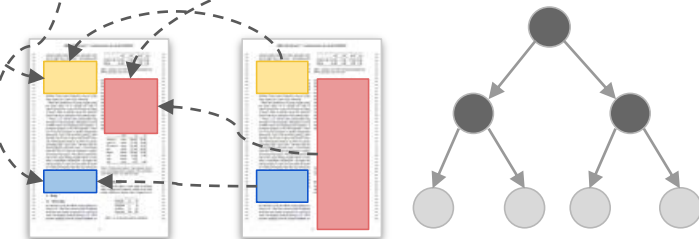
- Peer reviewing domain
- Three annotation layers
 - Linking
 - Versioning
 - Pragmatic tagging
- Explicit linker and version aligner
- Intertextual Graph library



Kuznetsov et al. 2022 “*Revise and Resubmit: An Intertextual Model of Text-based Collaboration in Peer Review*”.
Computational Linguistics 48(4).

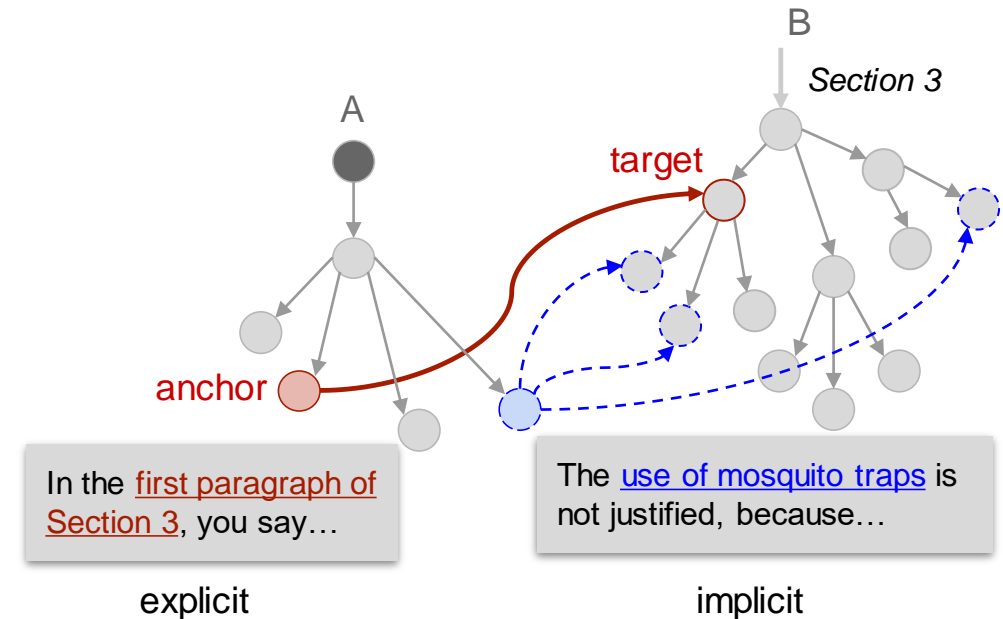


- (1) The paper describes a novel approach to <...>
- (2) It's well written and the idea is promising.
- (3) Perhaps you could include more details on time required to run the experiment.
- (4) I also found the Setup section a bit underspecified.



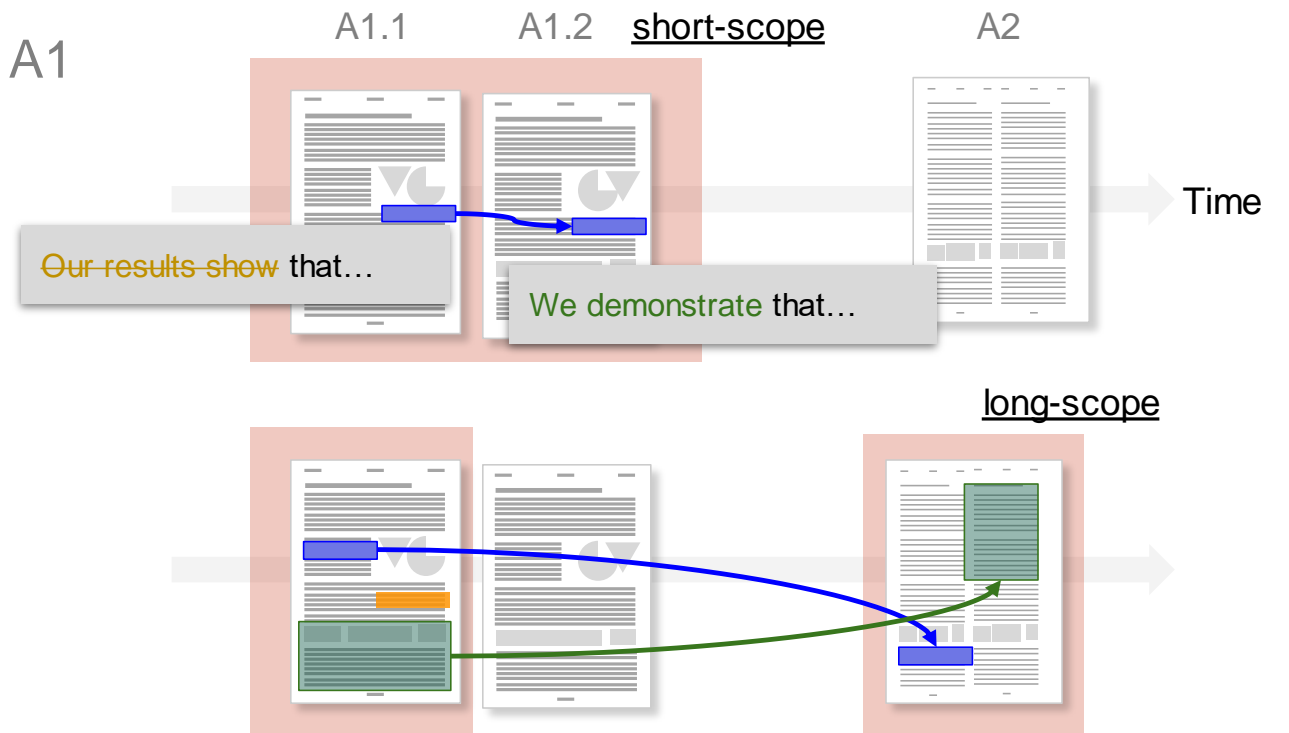
Linking: Text in Context

- Given document A that talks about document B
- Find anchors
parts of document A that talk about document B
- Find targets
parts of document B that the anchors refer to
- Generalization of
 - Citation span detection
 - Plagiarism detection
 - Evidence detection



Versioning: Text in Time

- Given document A2 which is a revision of document A1
- Find edits
correspondences between parts of A2 and A1
- Classify the edits
- Generalization of
 - Wikipedia edit analysis
 - Student essay revision analysis



Pragmatic tagging

- A peer review has a job → pragmatics
- Task: label sentences with general pragmatic tags

Recap	<i>The authors address the issue of...</i>
Weakness	<i>The discussion is superficial.</i>
Strength	<i>The paper is original and sound.</i>
Todo	<i>Please compare your method to...</i>
Other	<i>This idea reminded me of the work by...</i>
Structure	<i>Minor complaints:</i>

A very good attempt to present the Indian COVID-19 scenario by the authors. I congratulate them on their work. However a few queries:

Recap → - The data analysis has been performed on 1161 patients. To project it for such a large population has limited scope.

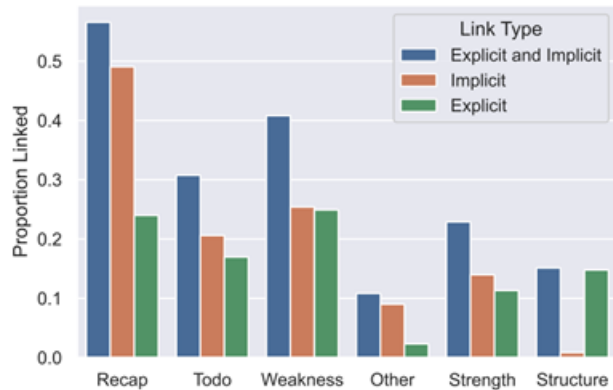
Other → - In COVID, most of the patients recover in due course. If possible, the SEIR model could have been used for a better picture.



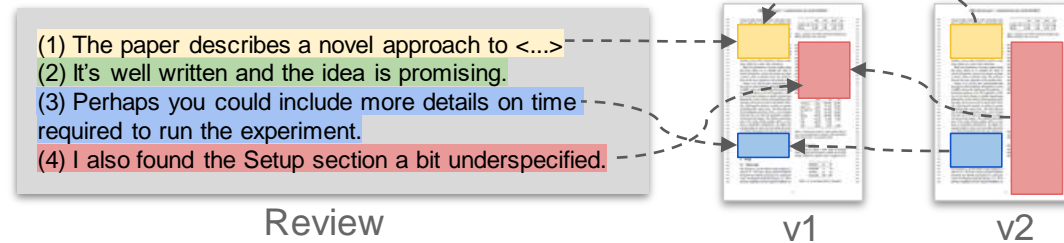
Dycke et al., 2023. *Overview of PragTag-2023: Low-Resource Multi-Domain Pragmatic Tagging of Peer Reviews (ArgMining-WS @ EMNLP)*

Pilot study in F1000RD

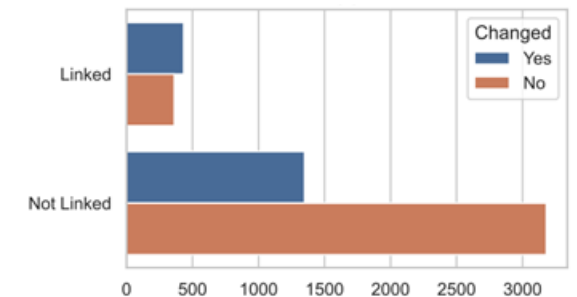
“What is the intention of this sentence, what is it about, and what change did it cause?”



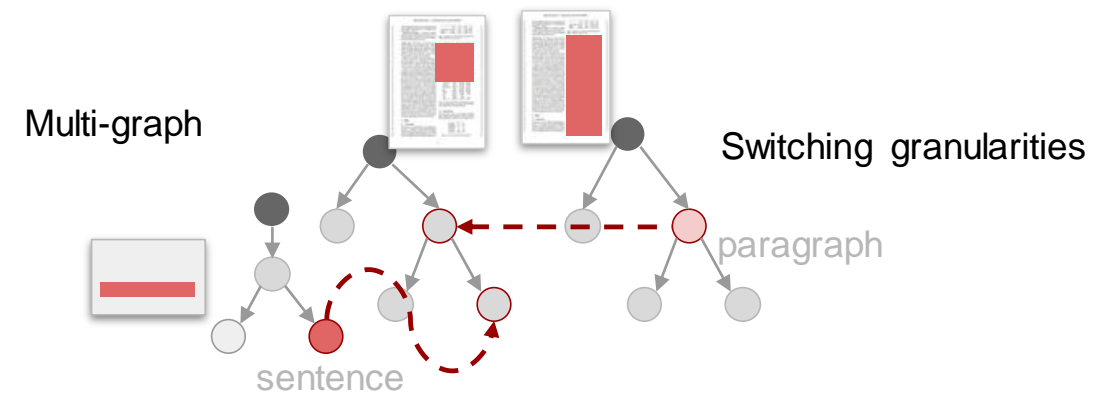
“What parts of papers receive most criticism?”



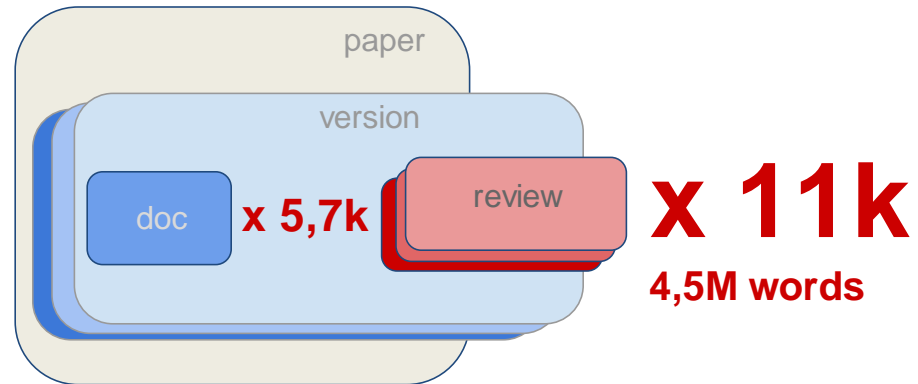
“Which types of comments trigger most change?”



Kuznetsov et al. 2022 “*Revise and Resubmit: An Intertextual Model of Text-based Collaboration in Peer Review*”. Computational Linguistics 48(4).



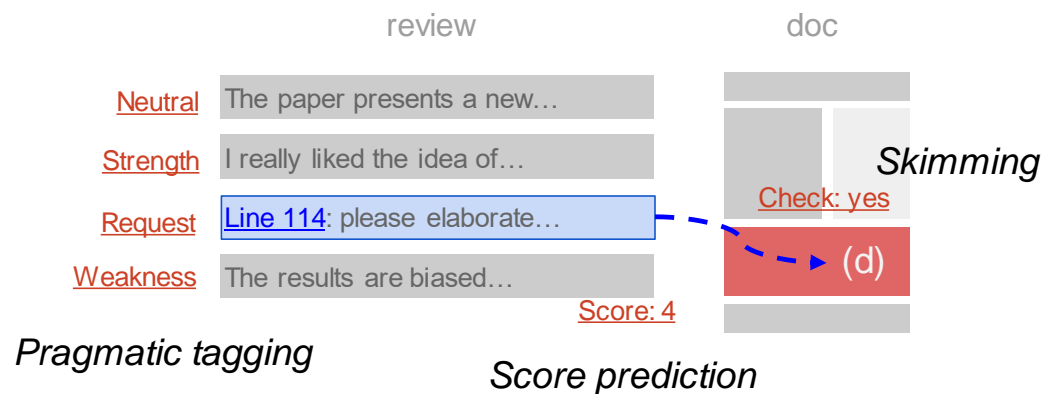
Getting the data: NLPEER



NLPEER

One-stop shop for NLP for peer review

- Large corpus
- Drafts, reviews and revisions
- Several domains and communities
- Open review, blind review
- Applied tasks



Dycke et al. 2022b “*NLPeer: A Unified Resource for the Computational Study of Peer Review*”. ACL 2023

NLPEER – Datasets

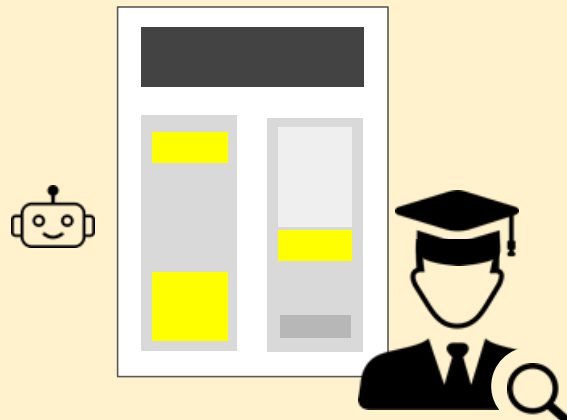
		domain	system	# papers	# reviews	# review tokens
NLPEER	ARR-22 NEW	CL/NLP	closed	476	684	266k
	COLING-20 NEW			89	112	45k
	ACL-17			136	272	100k
	CoNLL-16			22	39	16k
	F1000-22 NEW	multi	open	4949	10k	3.8M

NLPEER – Assisting Reviewers

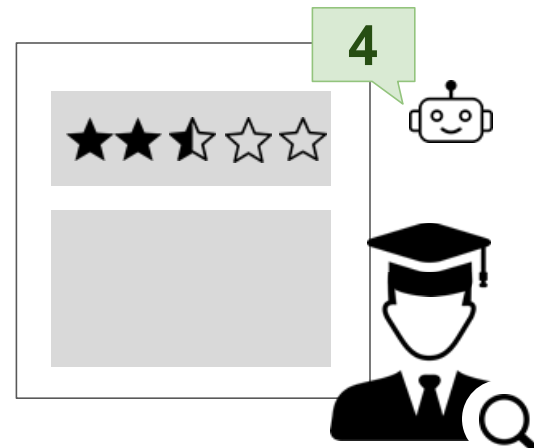
Exploiting the inter-textuality of peer reviews!

Guided Skimming for Peer Review

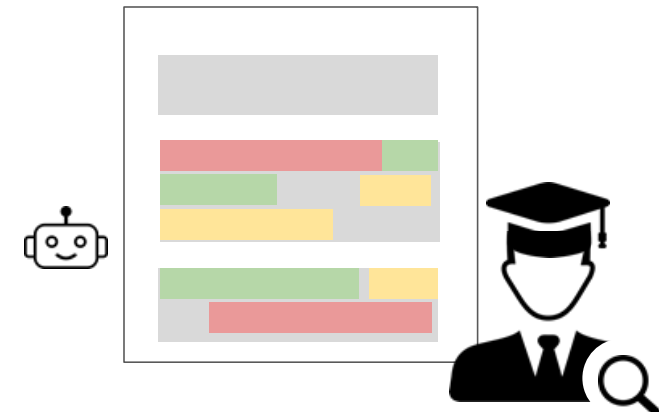
NEW



Review Score Prediction

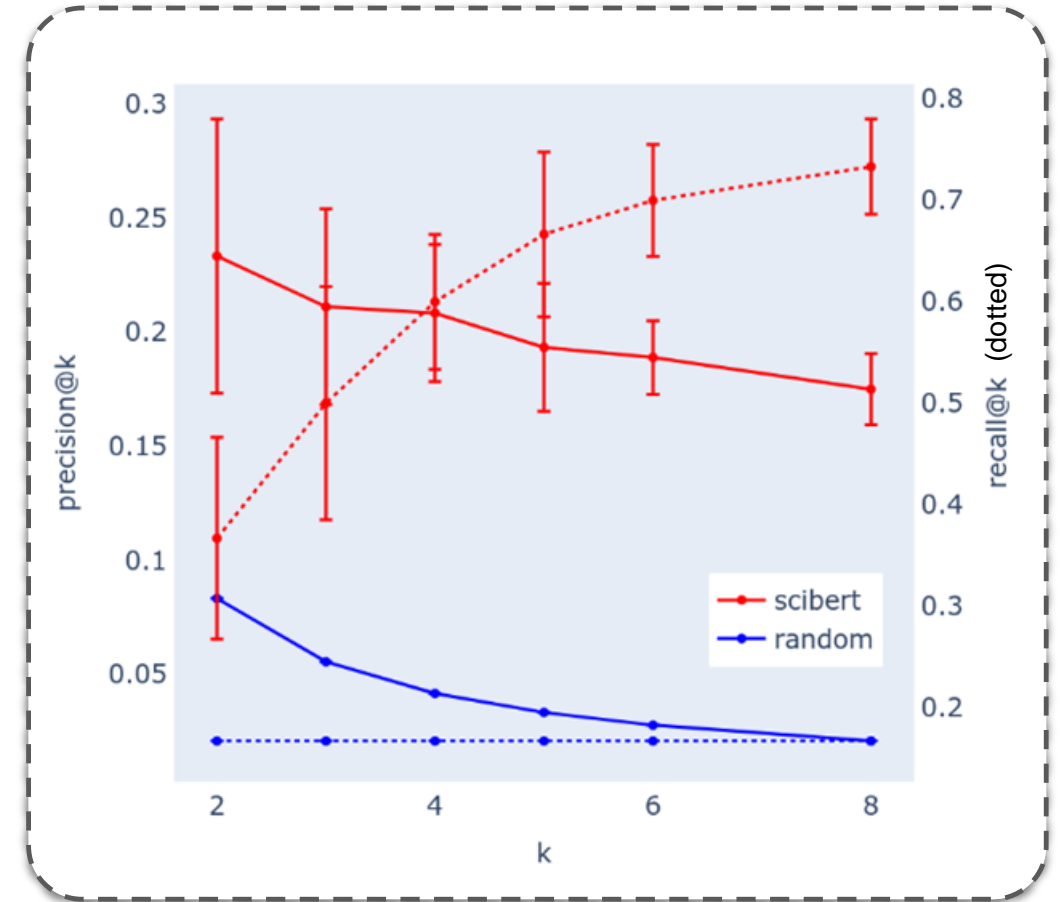
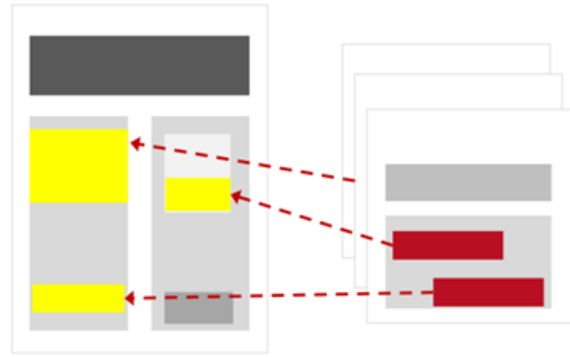


Pragmatic Labeling



NLPEER – Assisting Reviewers

Training

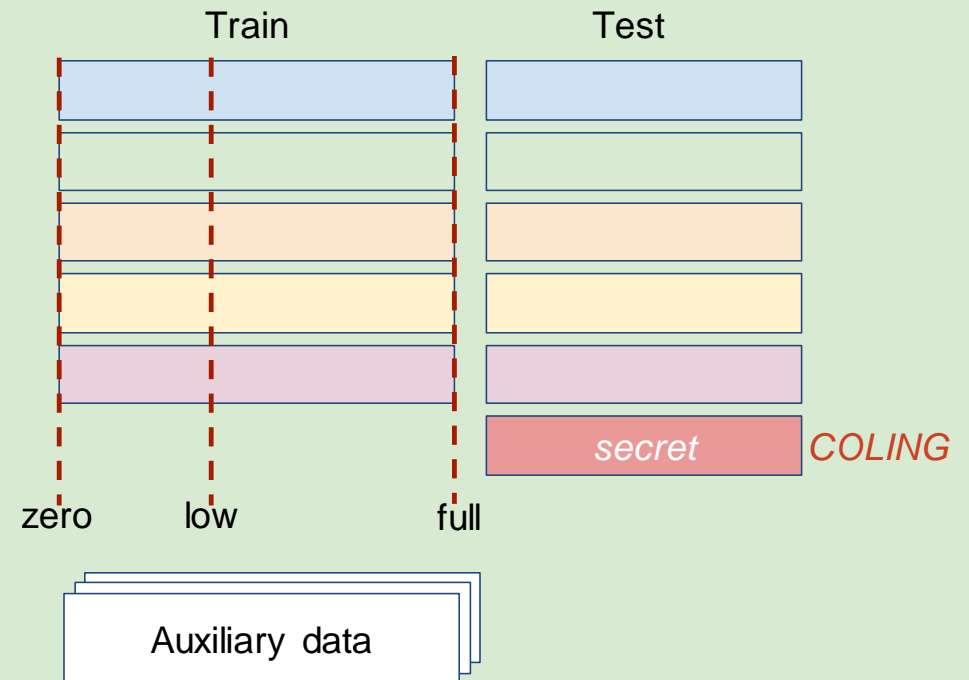


PragTag Shared Task @EMNLP-2023



- One task
 - Pragmatic tagging
- Five domains + secret
 - Disease outbreaks
 - Computational biology
 - Medical case studies
 - R Package development
 - Scientific policy
- Three data conditions
 - Zero-shot, low data, full data

“The paper omits crucial details about the data collection”
→ Weakness



PragTag Shared Task @EMNLP-2023



- 5 teams
 - CATALPA_NLP
 - DeepBlueAI
 - MILAB
 - NUS-IDS
 - SuryaKiran
- Baselines
 - Fine-tuned RoBERTa
 - Majority



PragTag Shared Task @EMNLP-2023



best, second-best

	team	<u>mean</u>	case	diso	iscb	rpkg	scip	secret
🏆	DeepBlueAI	<u>84.1</u>	82.9	84.1	82.8	<u>86.0</u>	<u>89.0</u>	<u>80.1</u>
🥈	NUS-IDS	83.2	83.8	<u>85.4</u>	<u>83.3</u>	84.8	87.8	74.1
	MILAB	82.4	<u>84.0</u>	83.7	80.1	85.4	86.5	74.9
	SuryaKiran	82.3	82.0	82.8	81.8	82.8	86.5	77.9
	CATALPA	81.3	80.8	82.0	81.1	82.5	82.5	78.8
🤝	Ensemble	84.4	84.0	85.2	83.3	87.3	88.7	78.0
	RoBERTa	80.3	80.3	80.8	79.9	83.1	83.8	73.7
	Majority	8.0	9.3	7.3	7.5	8.6	7.9	7.3

Different performance across F1000RD domains + a drop on secret data



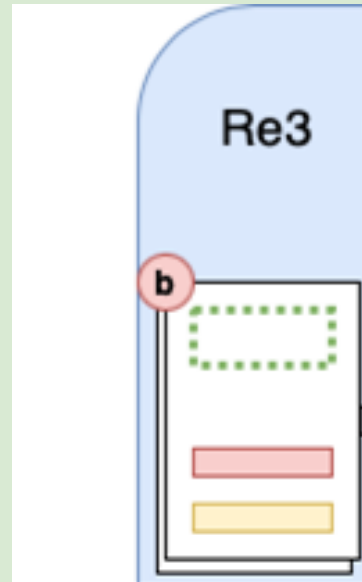
Nils Dycke, Ilya Kuznetsov, and Iryna Gurevych. 2023. *Overview of PragTag-2023: Low-Resource Multi-Domain Pragmatic Tagging of Peer Reviews*. In *Proceedings of the 10th Workshop on Argument Mining*, pages 187–196, Singapore. Association for Computational Linguistics.

Re3: A holistic framework for document revision



TECHNISCHE
UNIVERSITÄT
DARMSTADT

To appear in ACL-2024

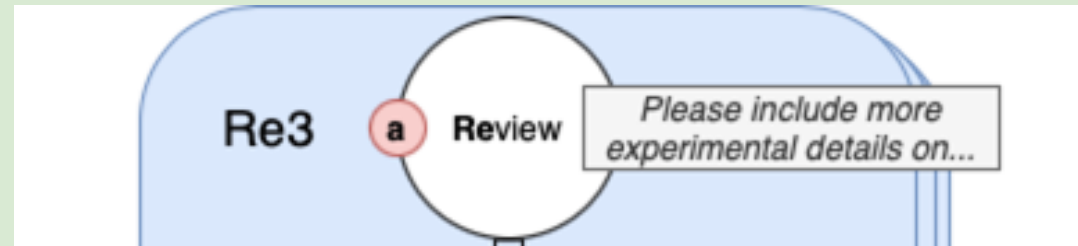


Re3: A holistic framework for document revision

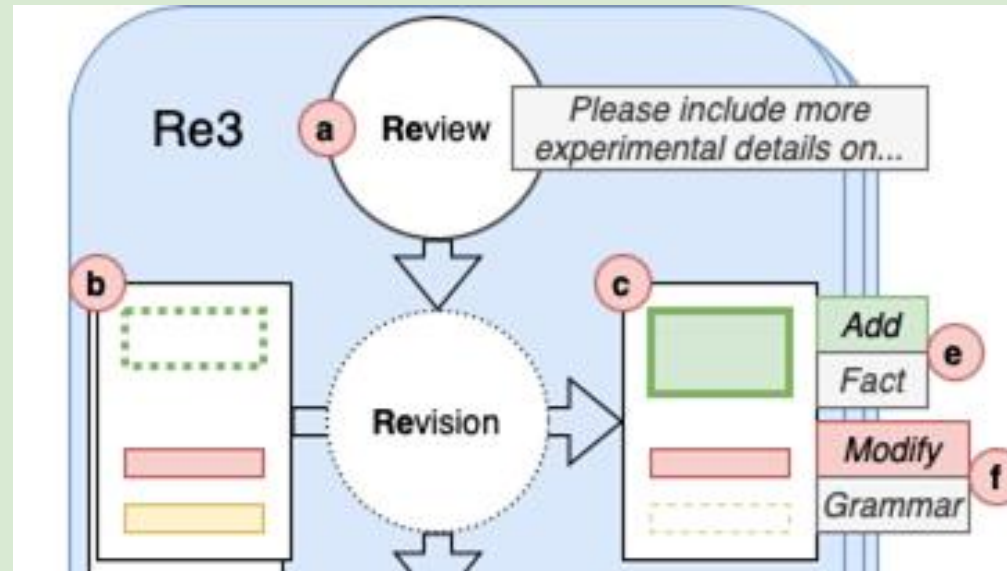


TECHNISCHE
UNIVERSITÄT
DARMSTADT

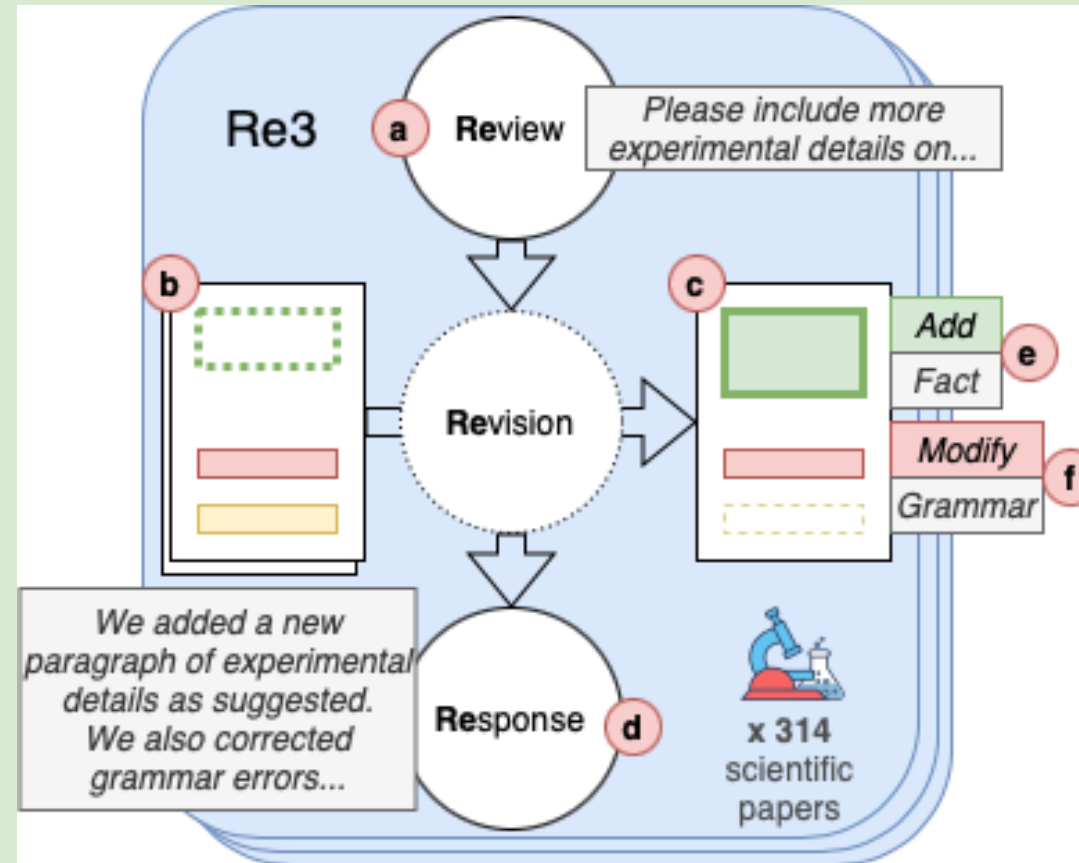
To appear in ACL-2024



Re3: A holistic framework for document revision



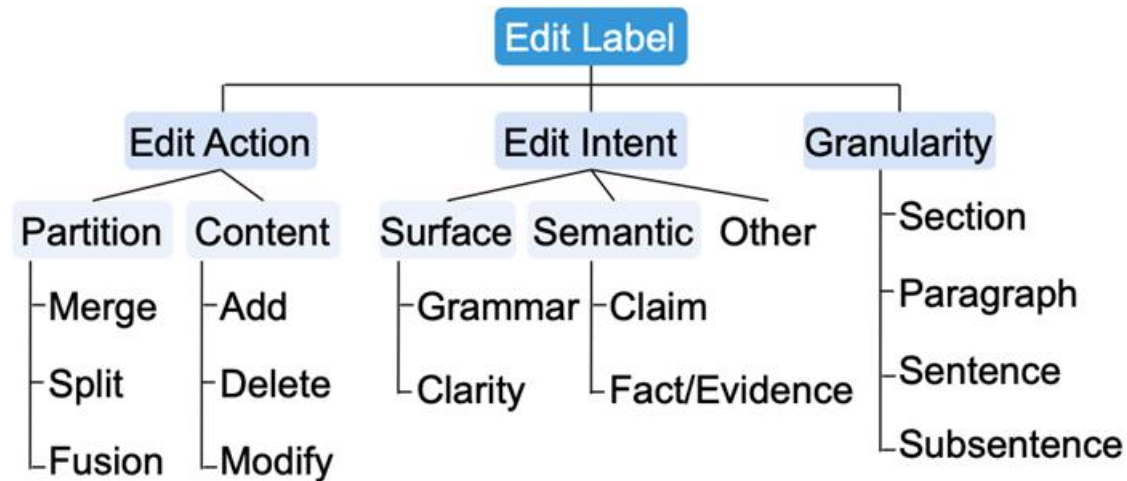
Re3: A holistic framework for document revision



Re3: A holistic framework for document revision



The screenshot shows a document editor with two document panes, 'doc1' and 'doc2'. A toolbar at the top includes icons for undo, redo, and other editing functions. A sidebar on the right is titled 'Layer' and contains a 'SentRelation' layer. Below this, there are 'From' and 'To' text input fields. A list of annotations is shown, including 'EditActionLabel' (Modify) and 'EditIntentionalLabel' (Grammar). A yellow box at the bottom right of the screenshot contains the text 'Labeling IAA 0.78 Krippendorff alpha'.



Re3: A holistic framework for document revision



- 314 full papers
from ARR and F1000Research
 - Parsed, unified, aligned, labeled
 - 11k+ labeled edits
- + reviews and revision requests
+ review-revision alignments
+ revision-edit summaries



Re3: A holistic framework for document revision



NLP tasks:

- Automatic revision alignment
- Edit intent classification
- Revision request extraction
- **Document edit summarization**



Qian Ruan, Ilya Kuznetsov, and Iryna Gurevych. 2024. Re3: A Holistic Framework and Dataset for Modeling Collaborative Document Revision. To appear in ACL-2024.

Re3: A holistic framework for document revision



NLP tasks:

- Automate
- Edit inter
- Revision
- Document

The revisions made to the long document can be summarized as follows:

1. Deletions:

- Several claims and statements regarding the testing and approval of new drugs, surgical procedures, and the comparison of procedures were deleted from the "Comments" section.

- A claim about the gender-based nature of the ban against Female Genital Cutting (FGC) and a suggestion for anti-FGC groups to advise on making FGC as safe as male circumcision were removed from the "Religious and cultural views" section.

- A statement about the unreliability of research based on reported data and individual memory was deleted from the "Comments" section.

- A claim about the strongest evidence coming from randomized controlled trials and cohort studies was removed from the "Comments" section.

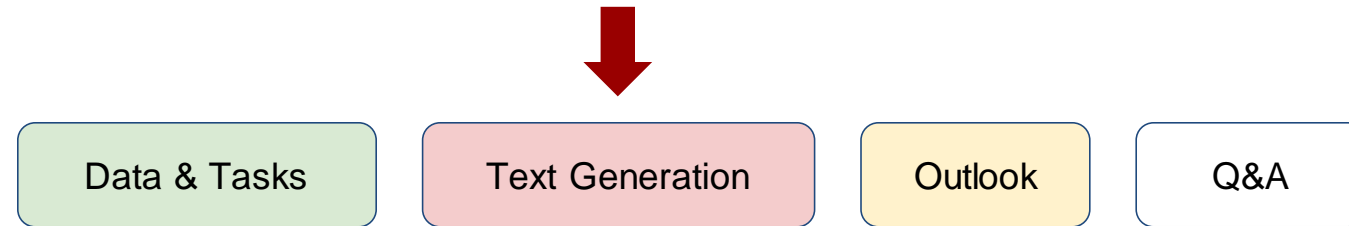
GPT-4

	#S	#W	Factuality	Comprehensiveness	Specificity	Compactness	Organization
human	19	346	100%	98.82%	95.56%	1.74	100% section
GPT-4	16	309	95.96%	79.09%	89.82%	2.36	72.5% action, 17.5% section

support the ban against FGC
rolled trial (RCT) to address

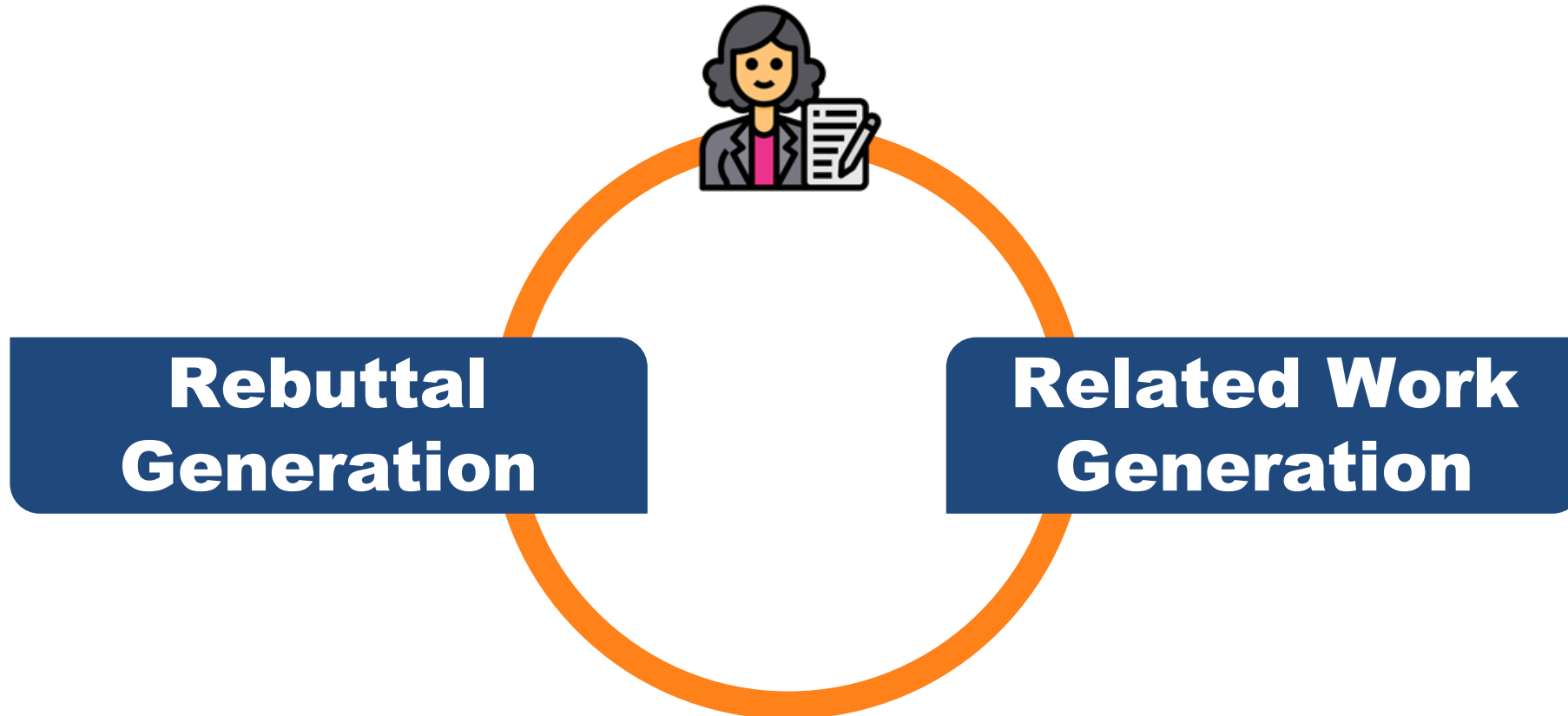


Qian Ruan, Ilya Kuznetsov, and Iryna Gurevych. 2024. Re3: A Holistic Framework and Dataset for Modeling Collaborative Document Revision. To appear in ACL-2024.



Text Generation

Assisting authors...



Rebuttal Generation



TECHNISCHE
UNIVERSITÄT
DARMSTADT

**Exploring Jiu-Jitsu Argumentation for
Writing Peer Review Rebuttals**
(Purkayastha et al., 2023) @ **EMNLP'23**



Related Work Generation

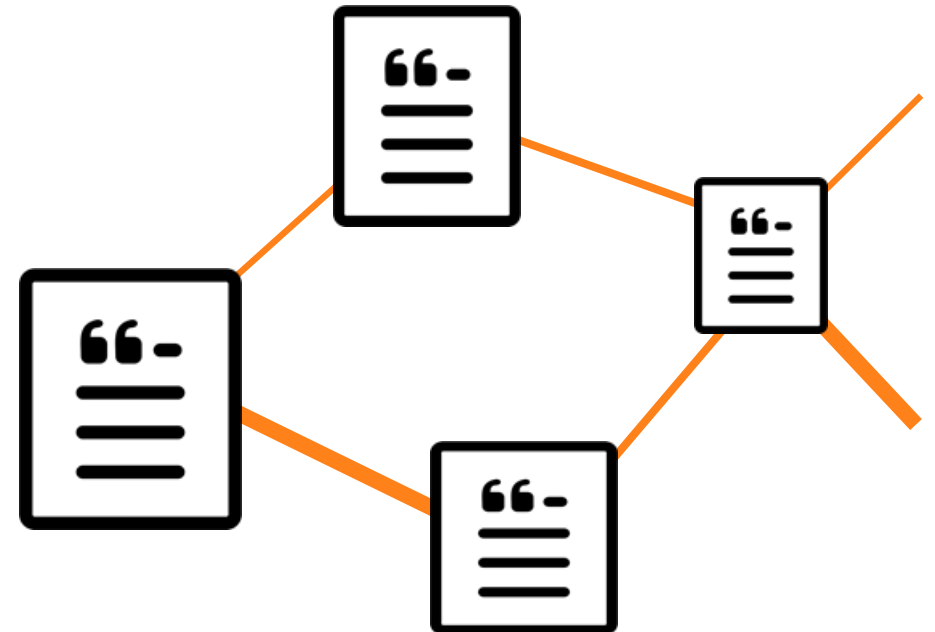


**CiteBench: A Benchmark for Scientific
Citation Text Generation**
(Funkquist et al., 2023) @ **EMNLP'23**

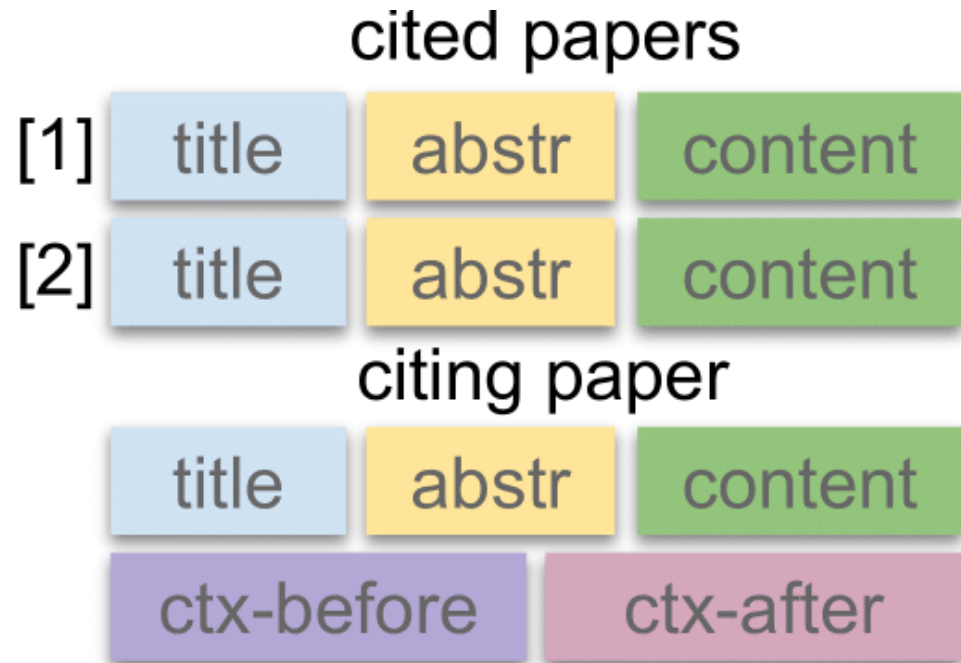


Assisting authors to define their related work...

- Research builds upon **prior publications**
- The publishing rates are **increasing**
- **Automated related work analysis** can
 - 🕒 Reduce time and effort
 - 🏆 Increase research quality



Citation Text Generation



→ generate:

Prior work has shown effective transfer from supervised tasks with large datasets, such as natural language inference [1] and machine translation [2].

Widely studied, but lacks unification!

CiteBench



TECHNISCHE
UNIVERSITÄT
DARMSTADT

**Standardized
baselines
+ comparison**



**Plug & Play
Evaluation kit**



**>300k samples
from 5 different
sources**



Unification of
 task definitions
 prior datasets

CiteBench 

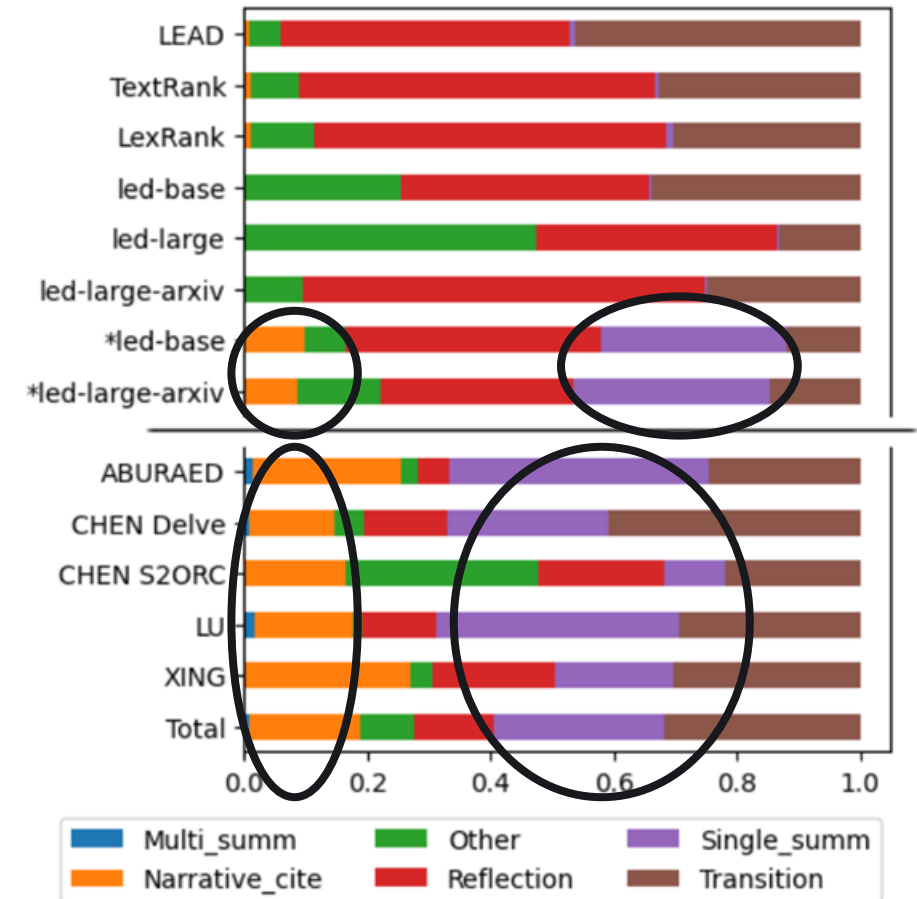
Performance

Model	ABURAED		CHEN Delve		CHEN S20RC		LU		XING	
	R-L	BertS	R-L	BertS	R-L	BertS	R-L	BertS	R-L	BertS
LEAD	11.32	75.42	11.48	74.70	11.34	74.33	13.34	75.89	10.55	75.25
TextRank	9.35	64.59	14.04	76.41	12.82	74.91	12.75	72.61	6.61	45.97
LexRank	10.80	74.90	12.93	75.96	12.85	75.11	14.24	76.70	10.06	75.14
led-base	9.06	74.84	5.55	70.86	5.41	70.55	7.36	72.35	10.07	74.87
led-large	8.30	73.26	6.22	69.57	6.22	69.77	6.89	70.51	9.35	74.30
led-large-arxiv	10.22	75.01	13.37	76.02	12.89	75.45	14.41	76.65	10.23	75.08
*led-base	13.44 _(0.03)	78.75	15.93	78.32	15.94	78.72	15.95	79.32	13.58 _(0.01)	78.49
*led-large-arxiv	14.90	79.0	16.27	78.13	15.92	78.58	16.53	79.41	12.42 _(0.01)	77.57

- Supervised models perform best
 - Larger not always better
- Extractive baselines perform surprisingly well
- Correlation between the metrics

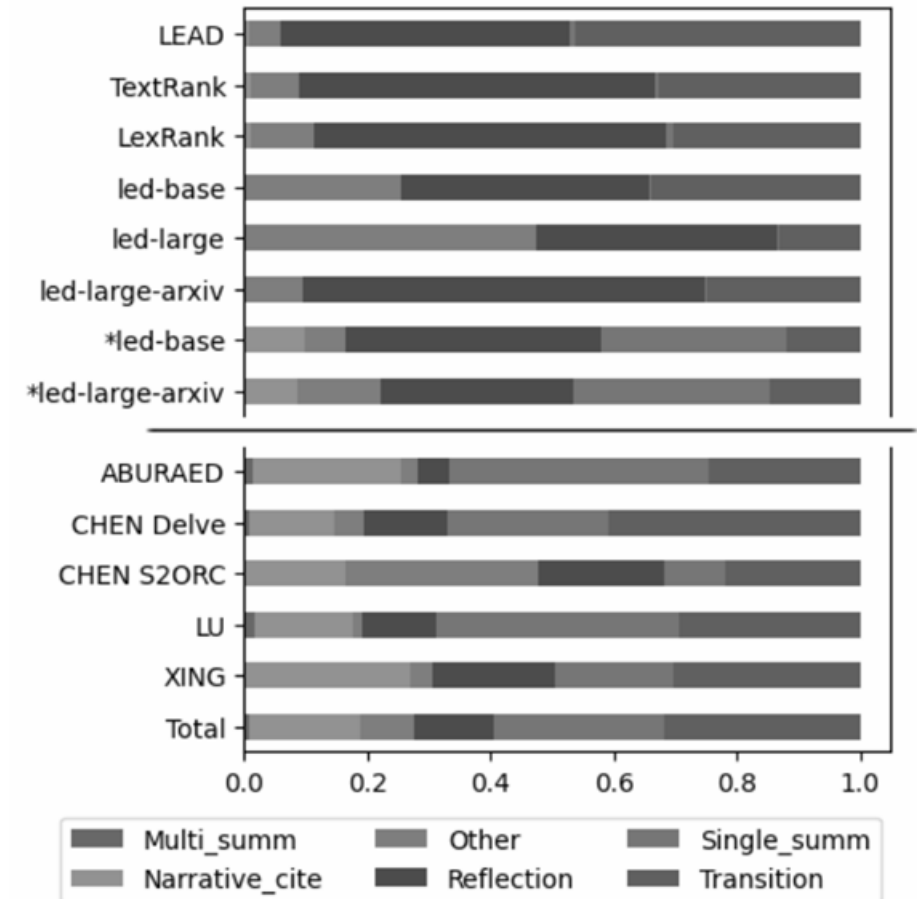
Findings

- Fine-tuned models perform best
- But there is still a room for improvement
- In particular, generated texts have a **different discourse structure** than human-written related work paragraphs →
- What information do we in fact need to generate accurate citation texts?



Findings

- Fine-tuned models perform best
- But there is still a room for improvement
- In particular, generated texts have a **different discourse structure** than human-written related work paragraphs →
- **What information do we in fact need to generate accurate citation texts?**



Systematic task exploration with LLMs



- What information is needed to generate citation texts?
- How to communicate this information to LLMs?
- How to measure the performance?

Key idea:

Due to flexible prompting and zero-shot capabilities LLMs allow easily experimenting with **alternative task definitions**

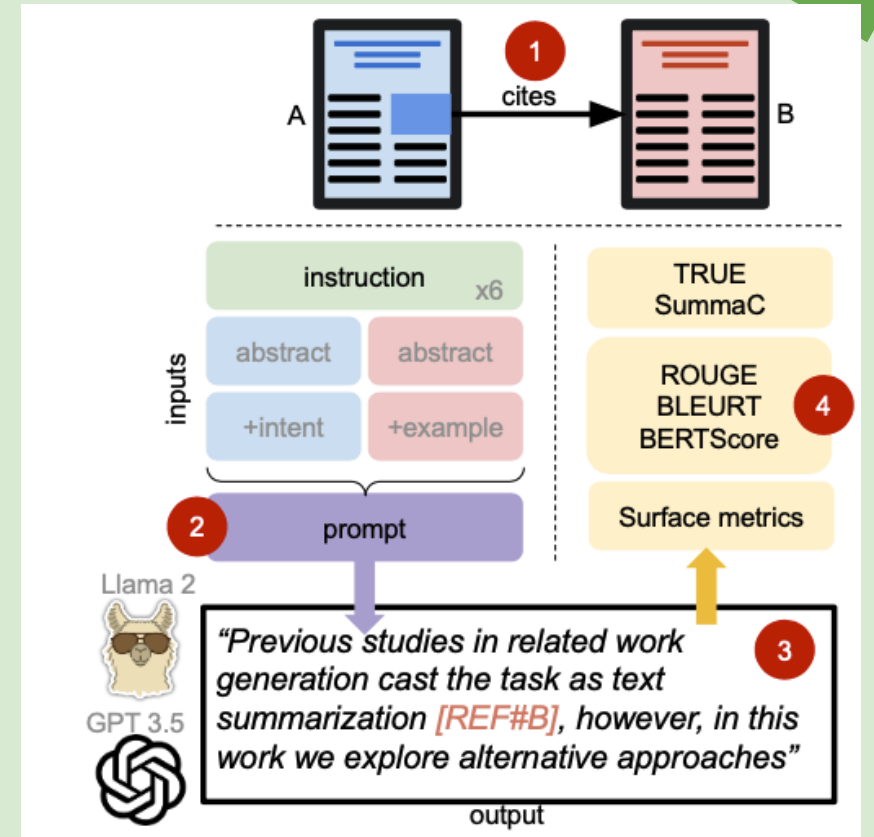
*Furkan Şahinuç, Ilya Kuznetsov, Yufang Hou, Iryna Gurevych.
2024. Systematic Task Exploration with LLMs: A Study in
Citation Text Generation. To appear in ACL-2024.*



Systematic task exploration with LLMs

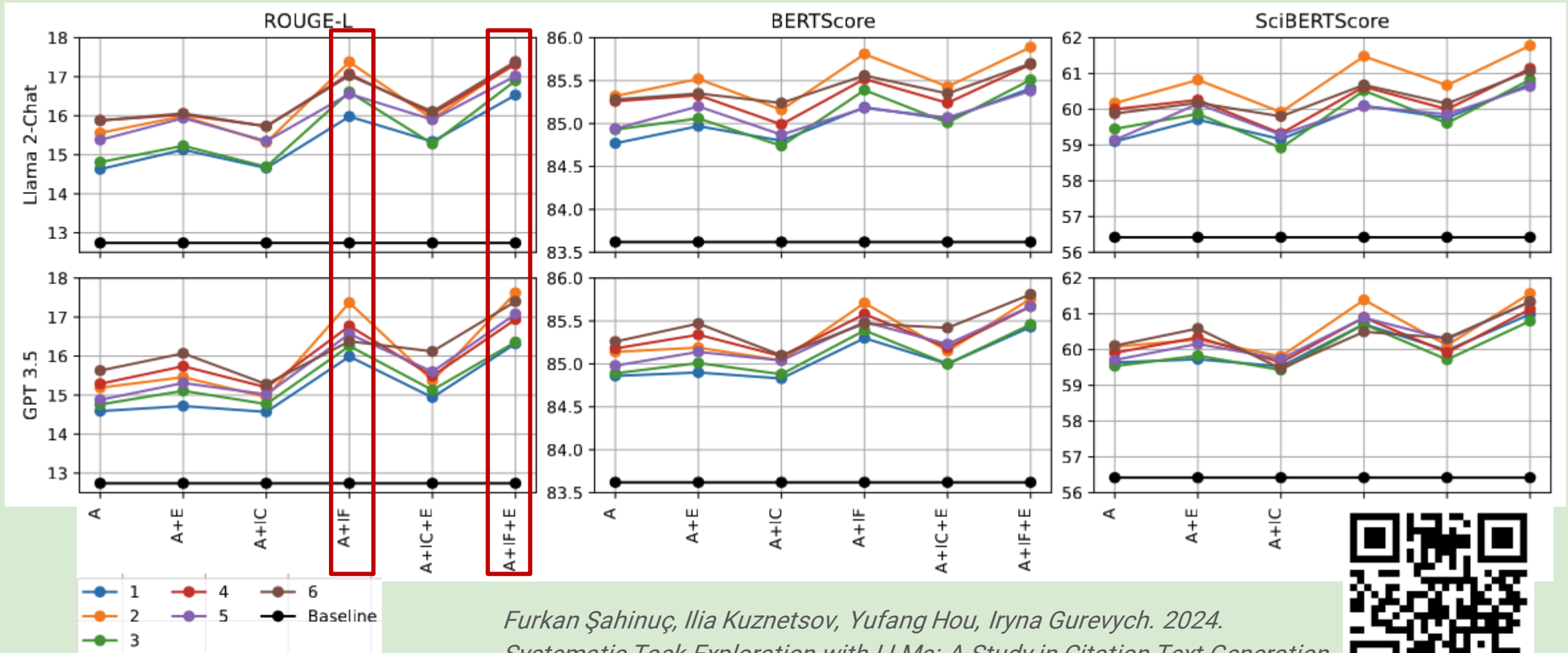


- LLMs for citation text generation
- **New dataset:** related work paragraphs from ACL
- **New framework:**
 - Prompt composition
 - Generation
 - Measurement
- Experiments on two models (LLaMA and GPT 3.5)
- Multiple NLG evaluation metrics
- Novel “free-form” citation intents*



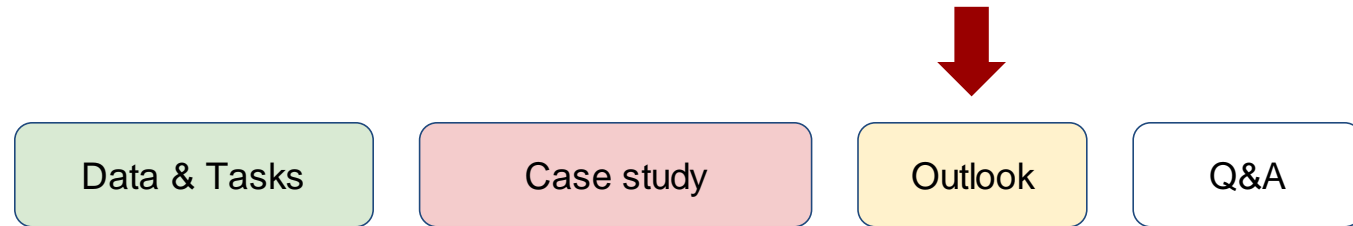
*see paper

Systematic task exploration with LLMs



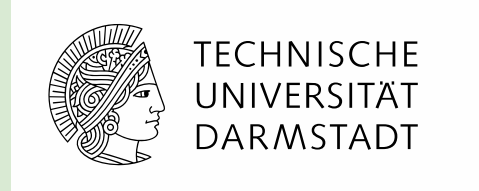
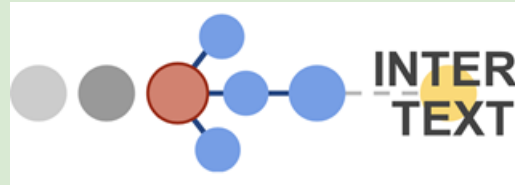
*Furkan Şahinuç, Iliya Kuznetsov, Yufang Hou, Iryna Gurevych. 2024.
Systematic Task Exploration with LLMs: A Study in Citation Text Generation.*





Outlook

Outlook



**Cross-Document
Implicit Linking**

**Human - AI
Collaborative Writing**



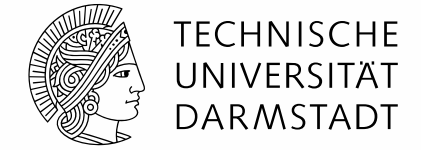
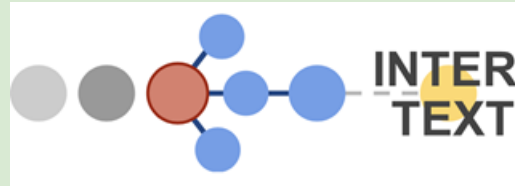
**Temporal Graphs
for the Science of
Science**

Novelty Assessment

A yellow five-pointed star icon located below the text "Novelty Assessment".

<https://intertext.ukp-lab.de/>

Challenges



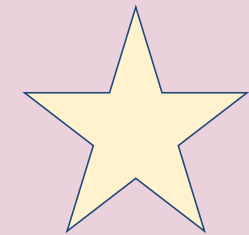
**Getting Data,
especially in
Europe**

**Technical challenges
- robustness, safety,
efficiency**



**Ethics and Dual-
Use**

**Human - AI
Collaboration**



<https://intertext.ukp-lab.de/>

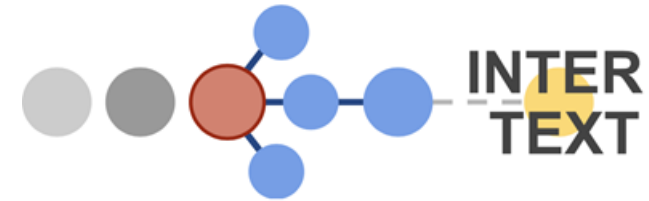
Exploring Our Work



TECHNISCHE
UNIVERSITÄT
DARMSTADT



intertext.ukp-lab.de/



Iryna Gurevych
Principal Investigator



Ilia Kuznetsov
Postdoc



Martin Tutek
Postdoc



Jan Buchmann
PhD Student



Nils Dycke
PhD Student



Max Eichler
PhD Student



Dennis Zyska
PhD Student



Qian Ruan
PhD Student

... and many more!