EACL 2026
MOROCCO
Palais Des Congres, Rabat
March 24 - 29, 2026

UBC
a place of mind
THE UNIVERSITY OF BRITISH COLUMBIA

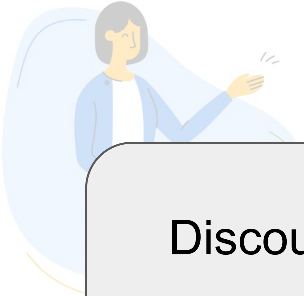
BeDiscovER: The Benchmark of Discourse Understanding in the Era of Reasoning Language Models

Chuyuan Li, Giuseppe Carenini

Department of Computer Science
The University of British Columbia

Why discourse understanding?

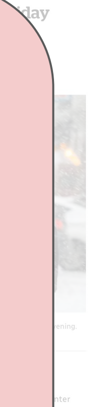
Speech



Dialogue

Scientific papers

News articles



Discourse is **a way of speaking.**

Rhetoric / Persuasion

But also:

- Text structure
- Communicative / behavior intentions
- Temporal reasoning
- Causal reasoning
- …

Yes, it can be eaten directly.

Why discourse understanding? – a few examples

Lexical

“recently”

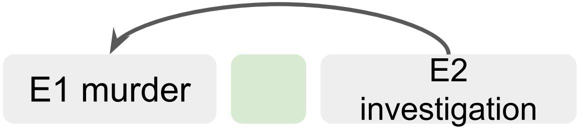
My brother ***just*** flew in to town.

“simply”

I ***just*** won’t stand for this injustice.

What is the sense of *just* in these contexts?

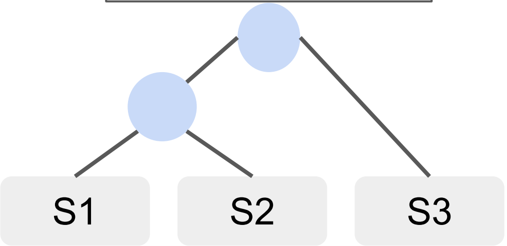
(Multi-)Sentence



E1 happens **before/after** E2?

E2 **explains/contradicts** E1?

Document



Correct **ordering** of S1, S2, S3?

How do S1, S2, S3 **interact** with each other?

Sheffield et al., Is it JUST semantics? a case study of discourse particle understanding in LLMs. Findings ACL 2025.

Why discourse understanding? – a few examples

Lexical

(Multi-)Sentence

Document

*Discourse understanding requires **lexical & semantic, temporal, rhetorical, commonsense**… knowledge.*

How well do modern LLMs understand discourse?

"recently"

My brother *just* flew i

"simply"

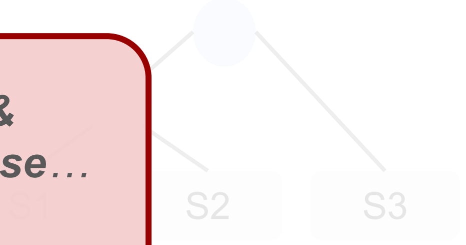
I *just* won't stand for this injustice.

Event 1

Event 2

E1 happens **before/after** E2?

E2 explains/contradicts E1?



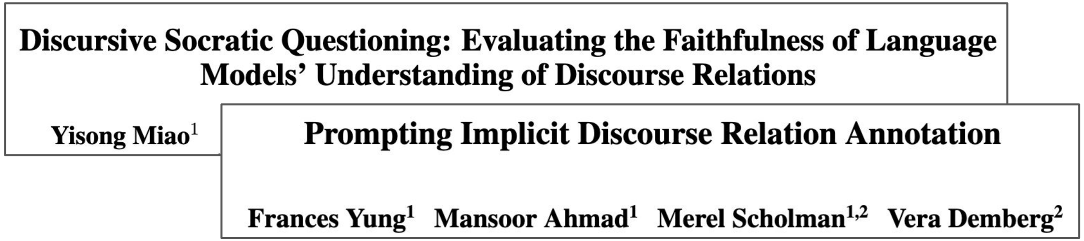
Correct **ordering** of S1, S2, S3?

How do S1, S2, … **eract** with each other?

Why discourse understanding? – existing studies

Theory oriented

PDTB-style relation classification



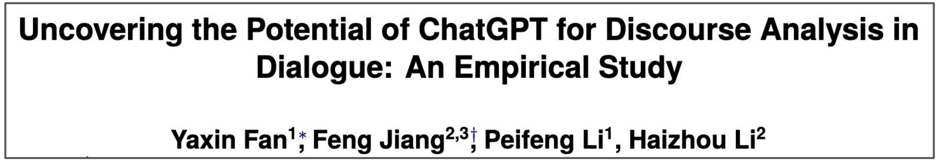
Discursive Socratic Questioning: Evaluating the Faithfulness of Language Models' Understanding of Discourse Relations

Yisong Miao[1]

Prompting Implicit Discourse Relation Annotation

Frances Yung[1] Mansoor Ahmad[1] Merel Scholman[1,2] Vera Demberg[2]

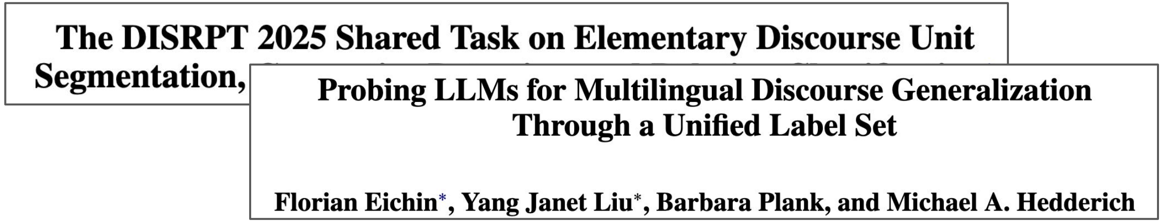
SDRT-style dialogue parsing



Uncovering the Potential of ChatGPT for Discourse Analysis in Dialogue: An Empirical Study

Yaxin Fan[1]*, Feng Jiang[2,3]†, Peifeng Li[1], Haizhou Li[2]

Generalization across frameworks



The DISRPT 2025 Shared Task on Elementary Discourse Unit Segmentation,

Probing LLMs for Multilingual Discourse Generalization Through a Unified Label Set

Florian Eichin*, Yang Janet Liu*, Barbara Plank, and Michael A. Hedderich

Why discourse understanding? – existing studies

Theory oriented

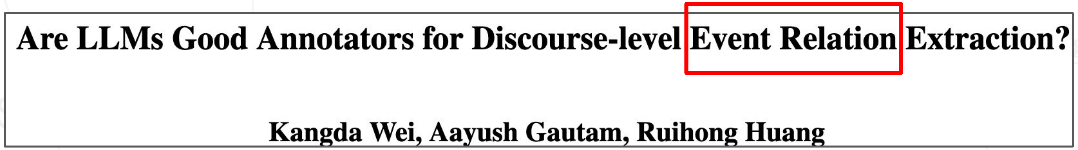
PDTB-style relation classification

Restricted on sentence-level, relation classification

Yisong Miao[1]

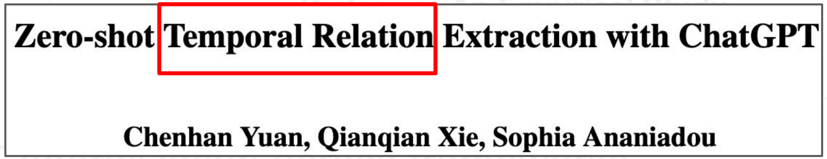
Prompting Implicit Discourse Relation Annotation

Task specific


Are LLMs Good Annotators for Discourse-level Event Relation Extraction?

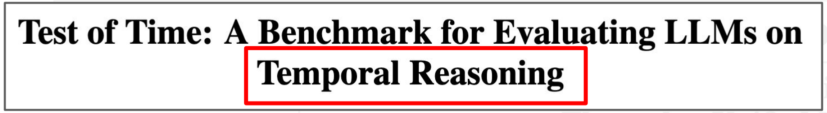
Kangda Wei, Aayush Gautam, Ruihong Huang

Uncovering the Potential of ChatGPT for Discourse Analysis in


Zero-shot Temporal Relation Extraction with ChatGPT

Chenhan Yuan, Qianqian Xie, Sophia Ananiadou

Generalization across frameworks


Test of Time: A Benchmark for Evaluating LLMs on Temporal Reasoning

Discourse Unit

Discourse Generalization Through a Unified Label Set

Florian Eichin*, Yang Janet Liu*, Barbara Plank, and Michael A. Hedderich

Why discourse understanding? – existing studies

Theory oriented → Restricted on sentence-level, relation classification

PDTB-style relation classification

Yisong Miao[1]

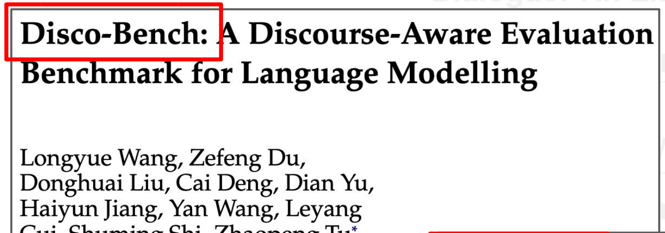
Prompting Implicit Discourse Relation Annotation

Frances Yung[1] Mansoor Ahmad[1] Merel Scholman[1,2] Vera Demberg[2]

Task specific → Focus on certain aspects of discourse understanding

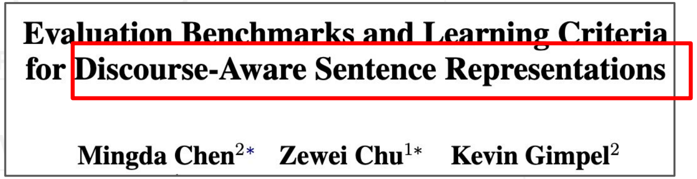
Uncovering the Potential of ChatGPT for Discourse Analysis in Dialogue: An Empirical Study

Discourse benchmarks



Disco-Bench: A Discourse-Aware Evaluation Benchmark for Language Modelling

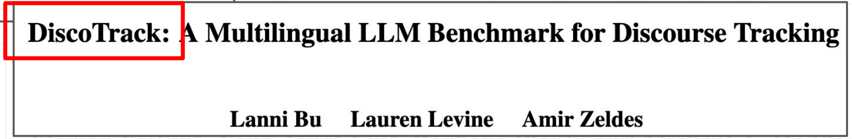
Longyue Wang, Zefeng Du, Donghuai Liu, Cai Deng, Dian Yu, Haiyun Jiang, Yan Wang, Leyang Cui, Shuming Shi, Zhaopeng Tu*



Evaluation Benchmarks and Learning Criteria for **Discourse-Aware Sentence Representations**

Mingda Chen[2*] Zewei Chu[1*] Kevin Gimpel[2]

...sk on Elementary Discourse Unit Segmentation,

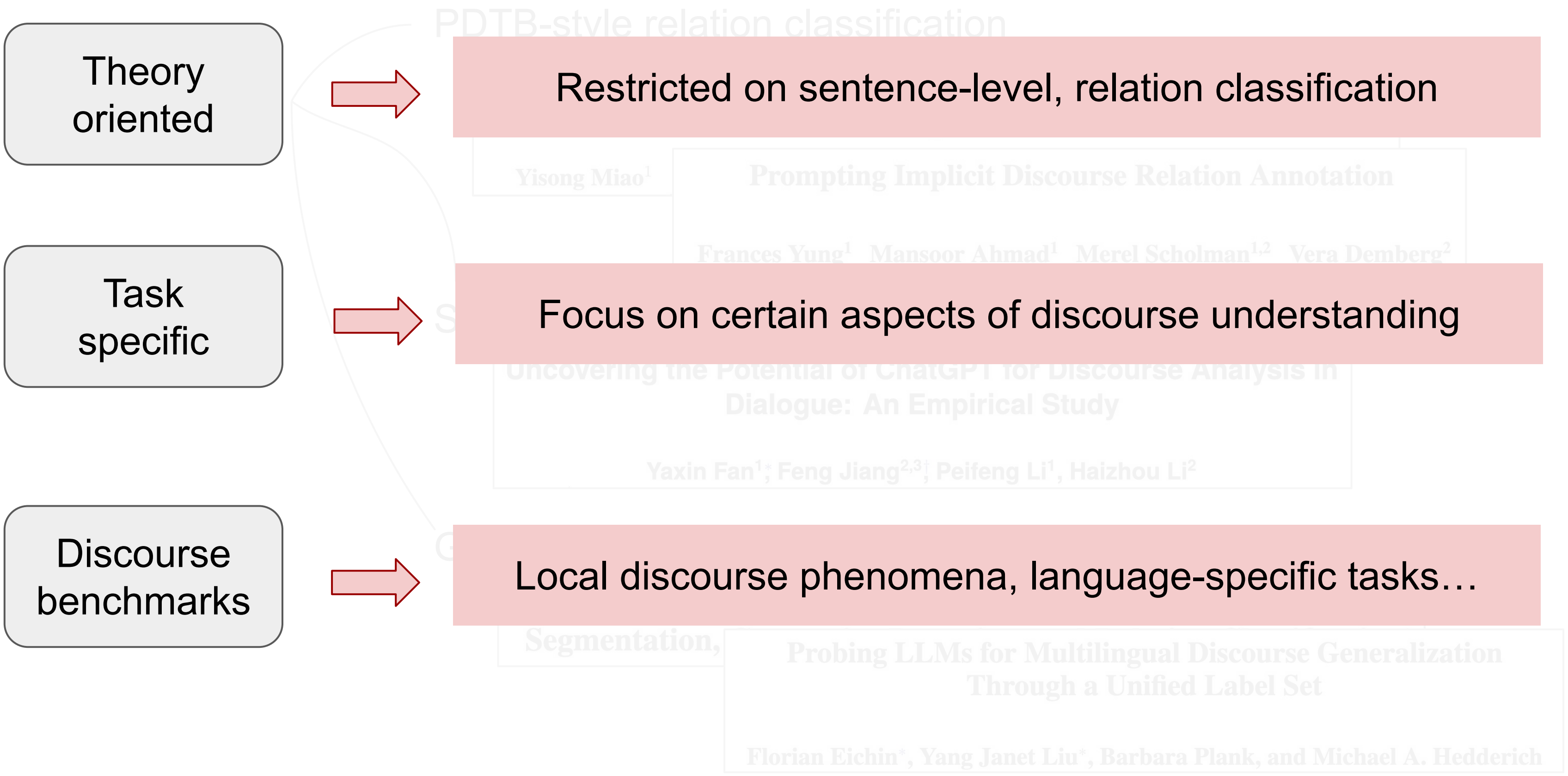


DiscoTrack: A Multilingual LLM Benchmark for Discourse Tracking

Lanni Bu Lauren Levine Amir Zeldes

Florian Eichin, Yang Janet Liu, Barbara Plank, and Michael A. Hedderich

Why discourse understanding? – existing studies



Theory oriented

Restricted on sentence-level, relation classification

Task specific

Focus on certain aspects of discourse understanding

Discourse benchmarks

Local discourse phenomena, language-specific tasks…

PDTB-style relation classification

Yisong Miao[1]

Prompting Implicit Discourse Relation Annotation

Frances Yung[1] Mansoor Ahmad[1] Merel Scholman[1,2] Vera Demberg[2]

Uncovering the Potential of ChatGPT for Discourse Analysis in Dialogue: An Empirical Study

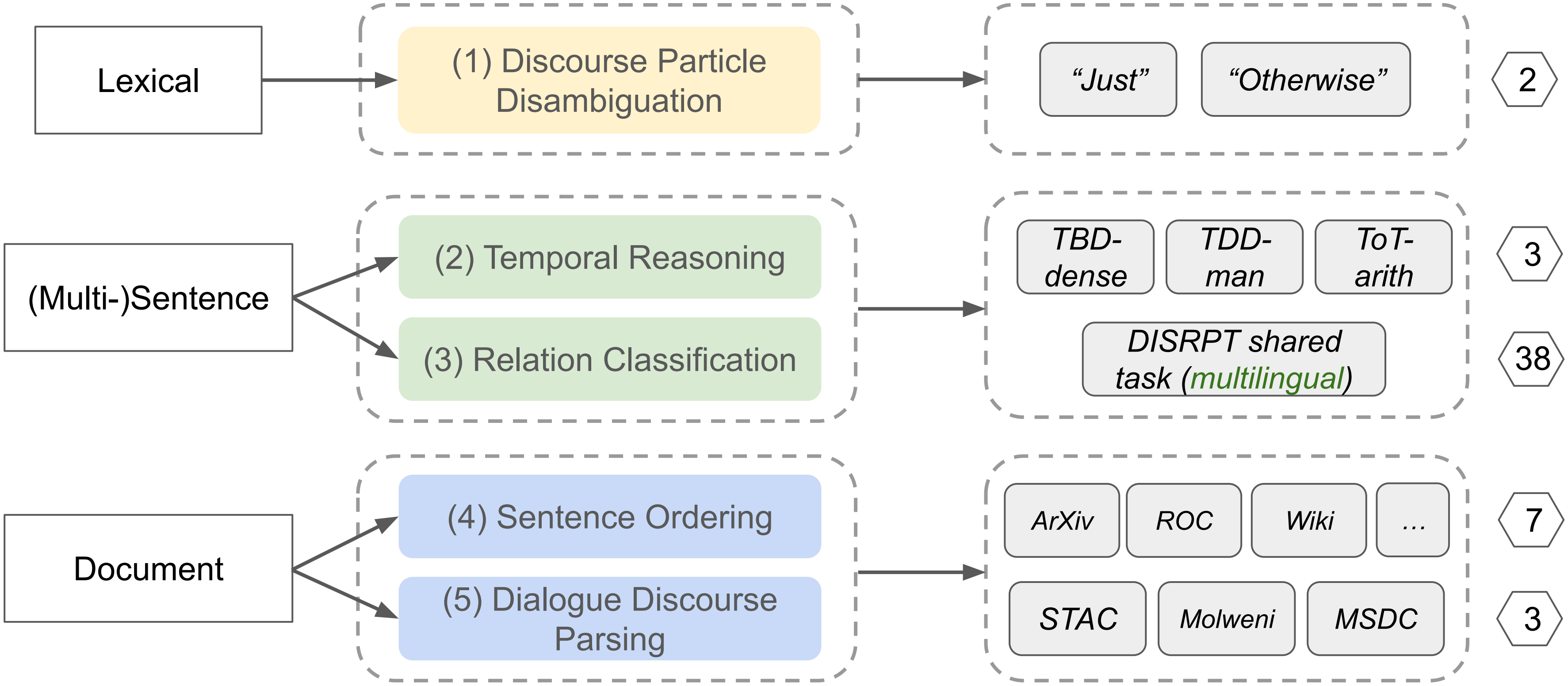
Yaxin Fan[1]*, Feng Jiang[2,3]†, Peifeng Li[1], Haizhou Li[2]

Segmentation,

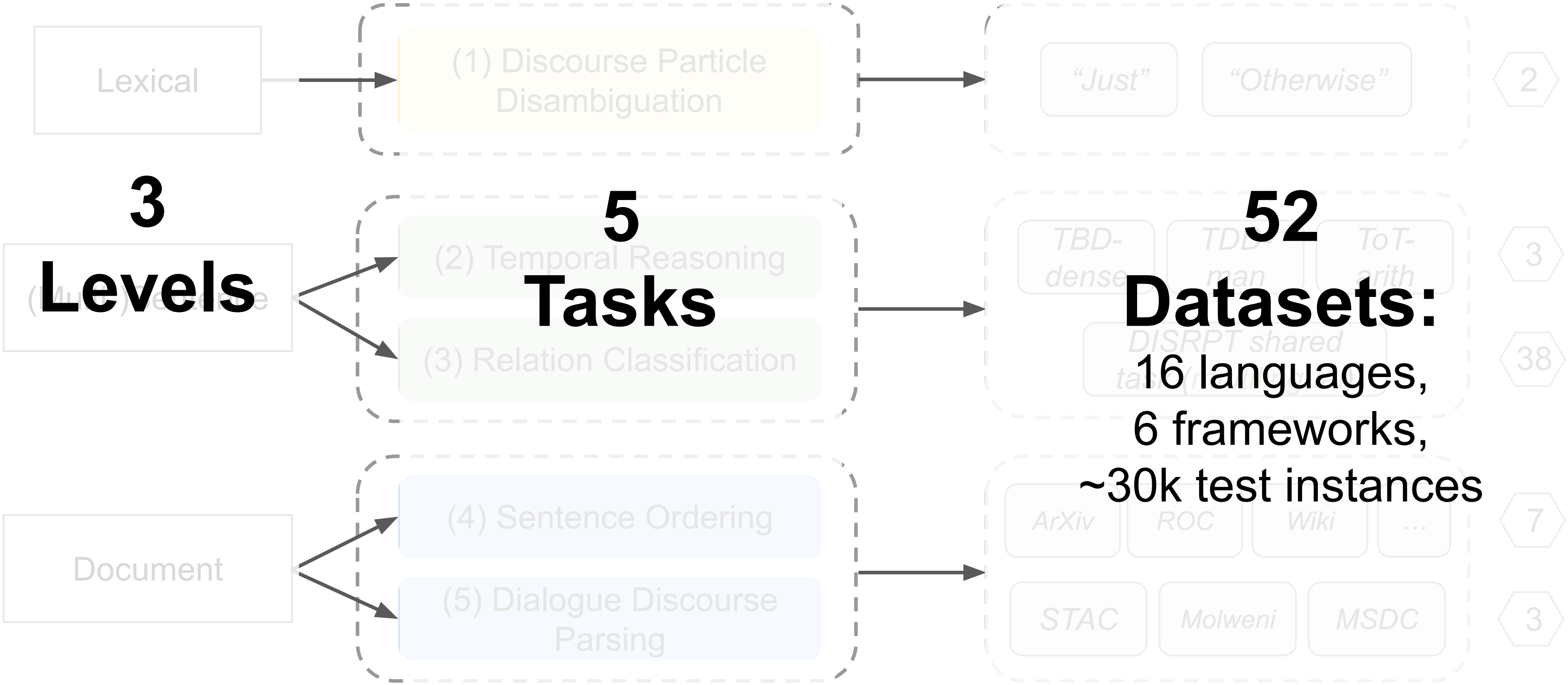
Probing LLMs for Multilingual Discourse Generalization Through a Unified Label Set

Florian Eichin*, Yang Janet Liu*, Barbara Plank, and Michael A. Hedderich

Our proposal: BeDiscovER!



Our proposal: BeDiscovER!



Lexical

(Multi-)Sentence

Document

(1) Discourse Particle Disambiguation

(2) Temporal Reasoning

(3) Relation Classification

(4) Sentence Ordering

(5) Dialogue Discourse Parsing

"Just" "Otherwise" 2

TBD-dense TDD-man ToT-arith 3

DISRPT shared task 38

ArXiv ROC Wiki … 7

STAC Molweni MSDC 3

3 Levels

5 Tasks

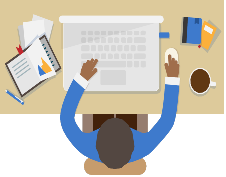
52 Datasets:

16 languages,
6 frameworks,
~30k test instances

BeDiscovER: Evaluation Setting

Open-ended Question-Answer Formatting

- **Unified evaluation pipeline**
- Classification tasks (1 2 3): fixed label space
- Parsing task (5): incremental generation task



Reasoning-oriented LLMs

GPT-5 Qwen3 DeepSeek-r1

Non reasoning-oriented LLMs

Llama-4

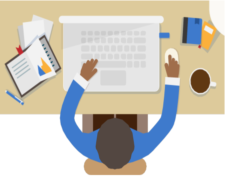
Qwen2.5
GPT4o

BeDiscovER: Evaluation Setting

System prompt:
… Choose one of the following six labels: **[Exclusionary, Unelaboratory, Unexplanatory, Emphatic, Temporal, Adjective].**

User prompt:
My brother ***just*** flew in to town.

Question: What is the function of the discourse marker "just" in the sentence above?



Reasoning-oriented LLMs

Temporal

GPT-5 Qwen3 DeepSeek-r1

Non reasoning-oriented LLMs

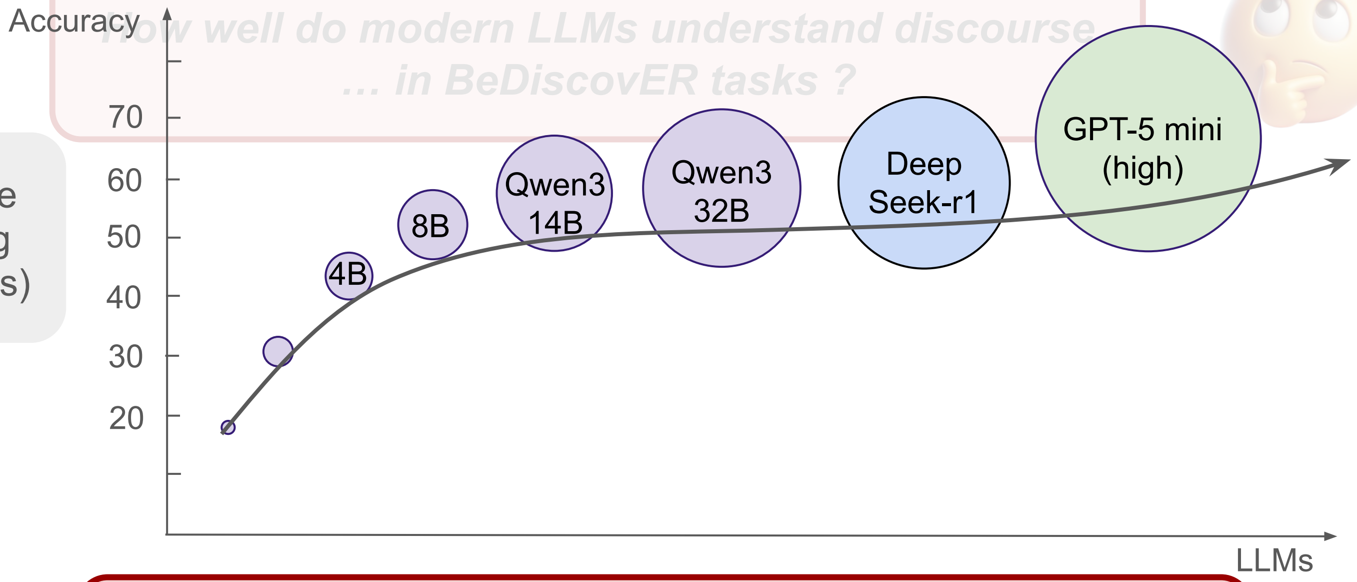
Llama-4

Qwen2.5
GPT4o

Performance: model scaling

How well do modern LLMs understand discourse … in BeDiscovER tasks ?

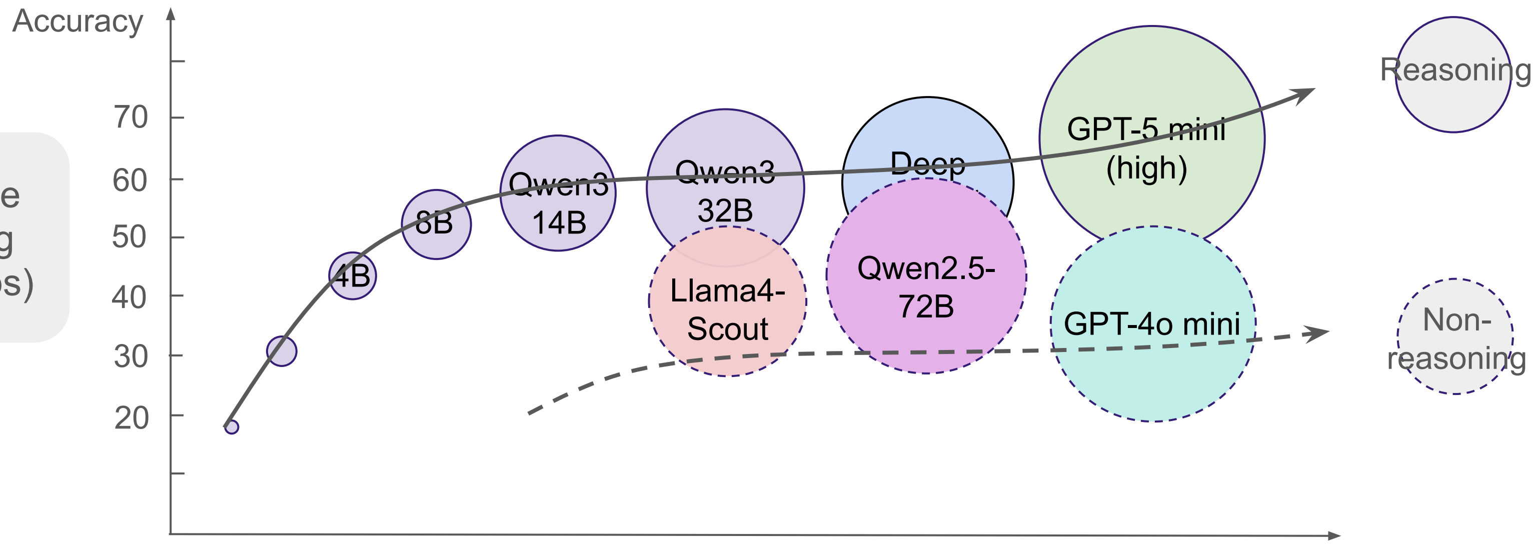
Sentence Ordering (ArXiv abs)



Bigger the model, better the performance – expected!

Performance: reasoning-oriented vs. non-reasoning LLMs

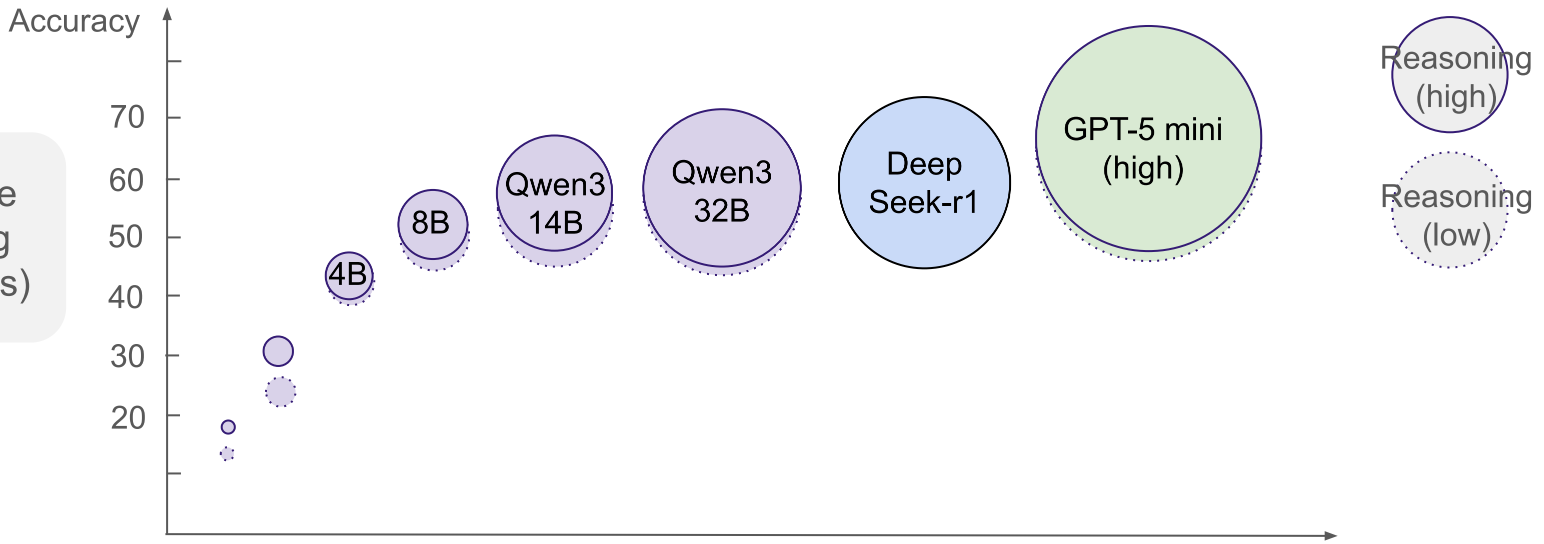
Sentence Ordering (ArXiv abs)



Reasoning-oriented LLMs outperform non-reasoning optimized LLMs – expected!

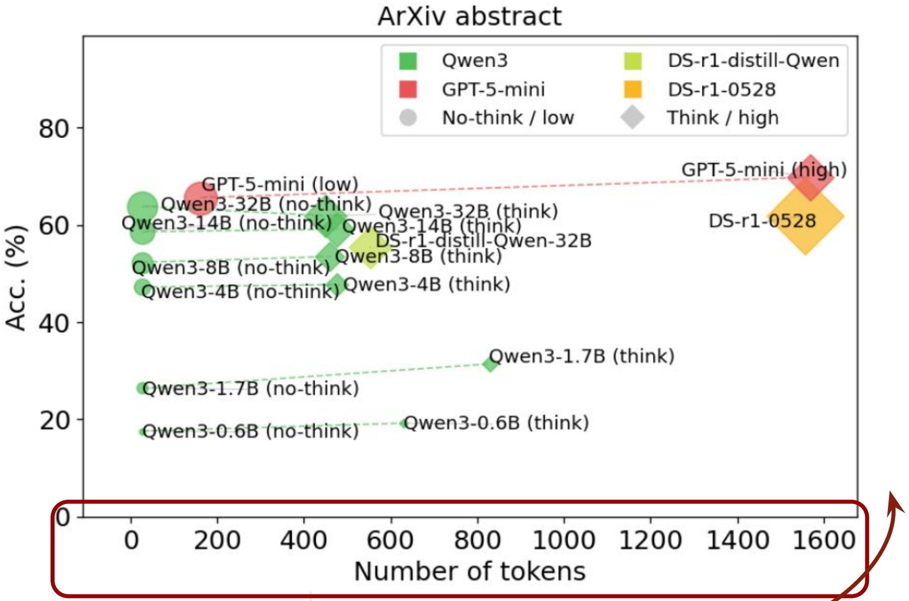
Performance: Higher reasoning effort, better result?

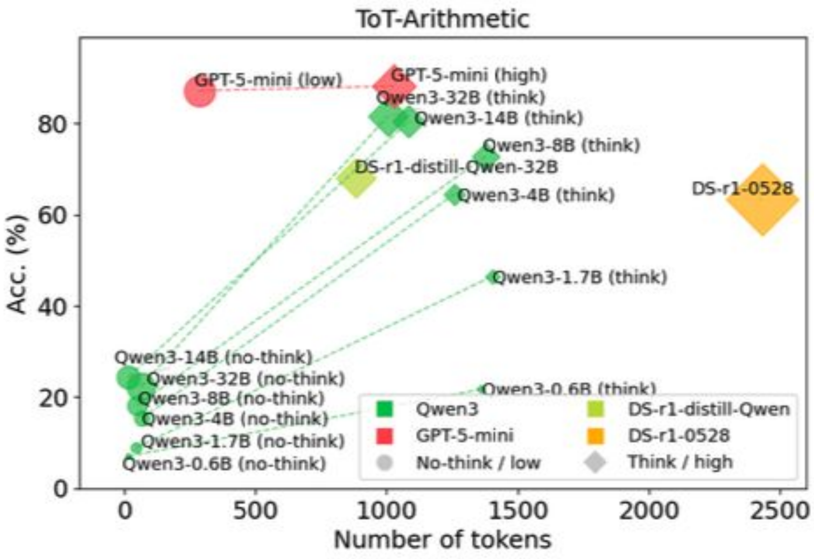
Sentence Ordering (ArXiv abs)



Higher thinking effort does not yield better outcome

Performance: Higher reasoning effort, better result?

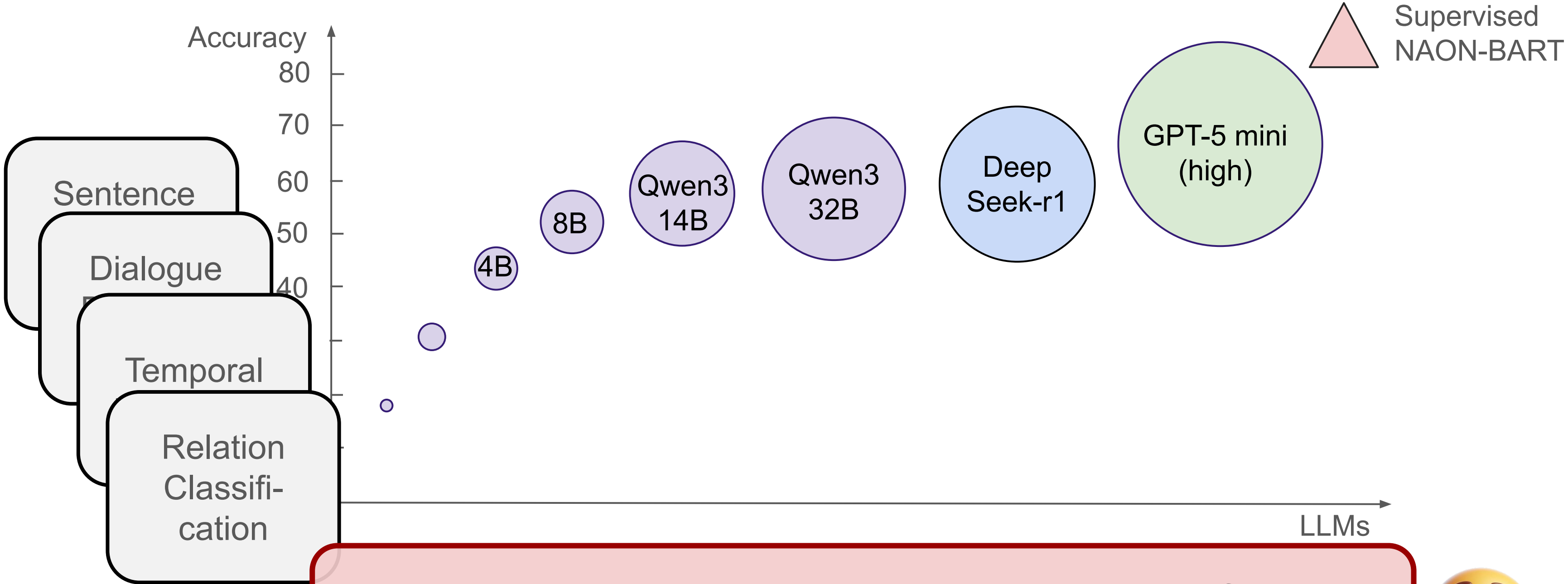




Models become verbose, yet, flat performance

*… except for **arithmetic** temporal: long reasoning lead to big gains!*

Performance: LLMs vs. supervised models

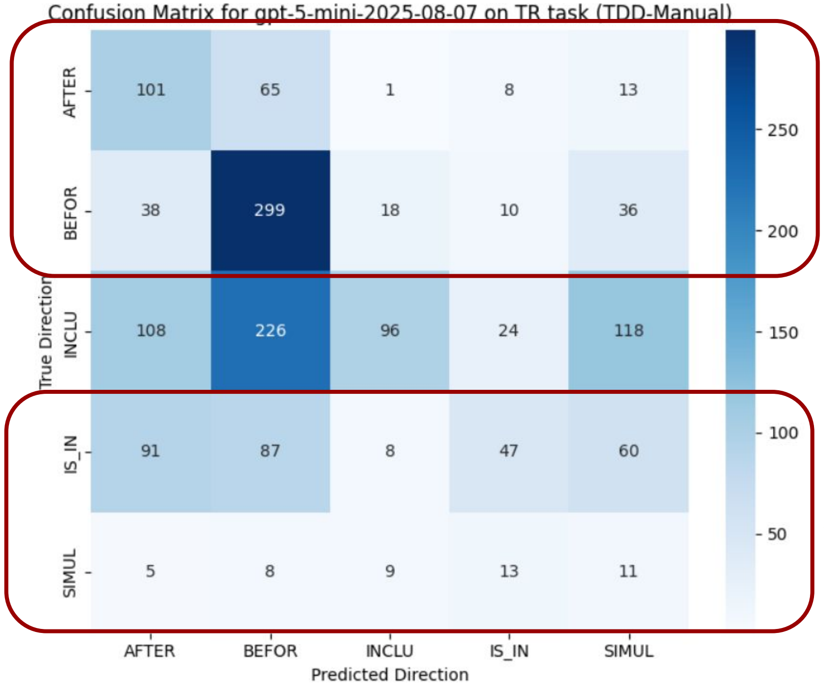


Reasoning-oriented LLMs show markedly lower performance (~10–30%) compared to supervised models.

Analysis: Fine-grained sense disambiguation

Temporal Reasoning

Relation Classifi-cation



Confusion Matrix for gpt-5-mini-2025-08-07 on TR task (TDD-Manual)

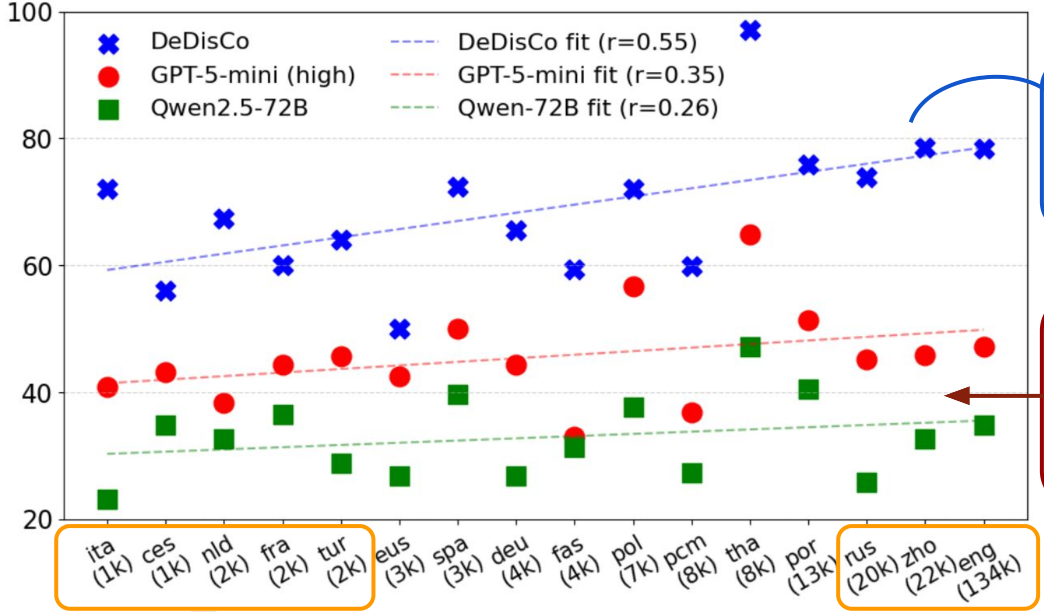
True Direction \ Predicted Direction	AFTER	BEFOR	INCLU	IS_IN	SIMUL
AFTER	101	65	1	8	13
BEFOR	38	299	18	10	36
INCLU	108	226	96	24	118
IS_IN	91	87	8	47	60
SIMUL	5	8	9	13	11

Identifies ***before/after*** relations

but fails to capture ***overlap*** or **containment**.

Analysis: Multilingual performance

Relation Classification



Supervised model show clear perform disparities (20%)

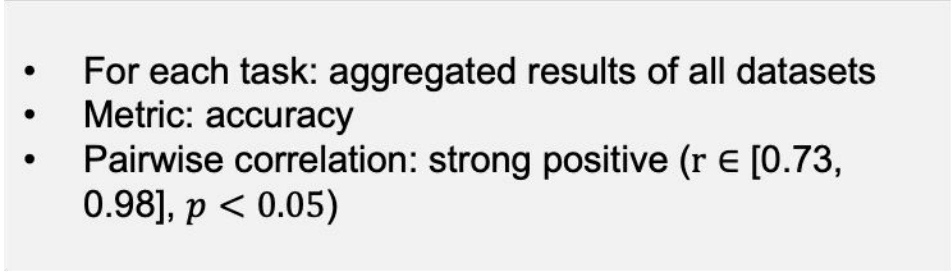
LLMs: nearly flat, lack of robust relation representation

low-resource languages

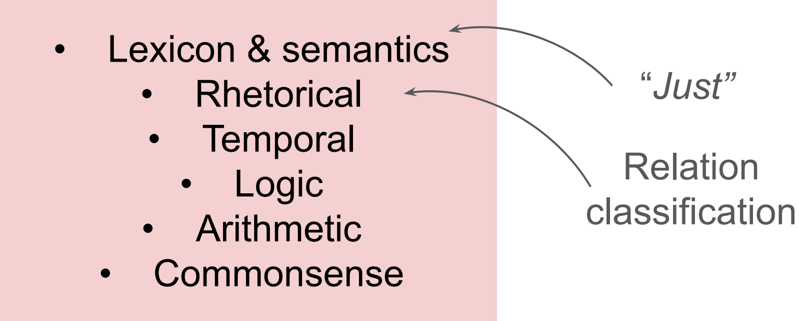
high-resource languages

Analysis: Inter-task correlation

Performance similarity (Pearson r)



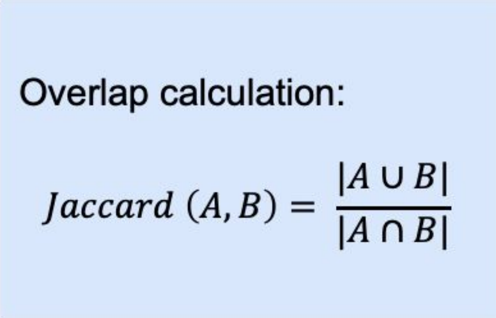
- For each task: aggregated results of all datasets
- Metric: accuracy
- Pairwise correlation: strong positive (r ∈ [0.73, 0.98], $p < 0.05$)



- Lexicon & semantics
- Rhetorical
- Temporal
- Logic
- Arithmetic
- Commonsense

"Just"

Relation classification

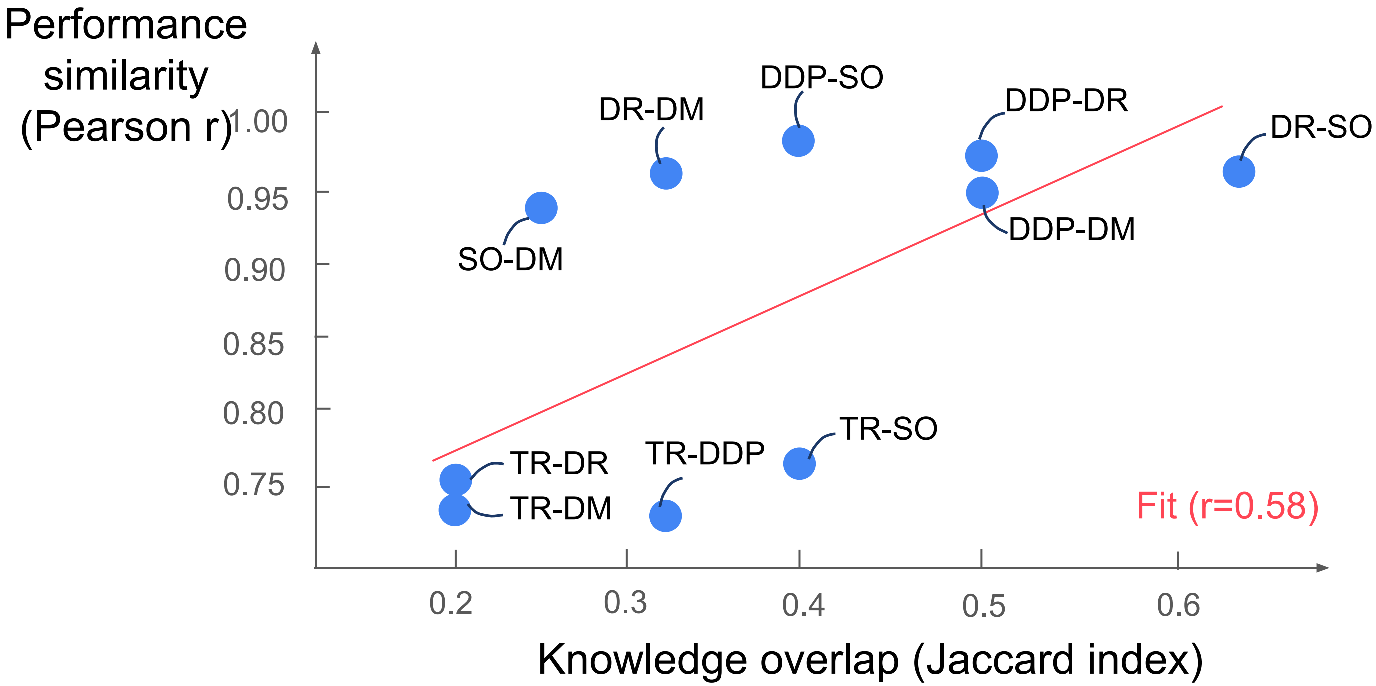


Overlap calculation:

$$Jaccard\ (A, B) = \frac{|A \cup B|}{|A \cap B|}$$

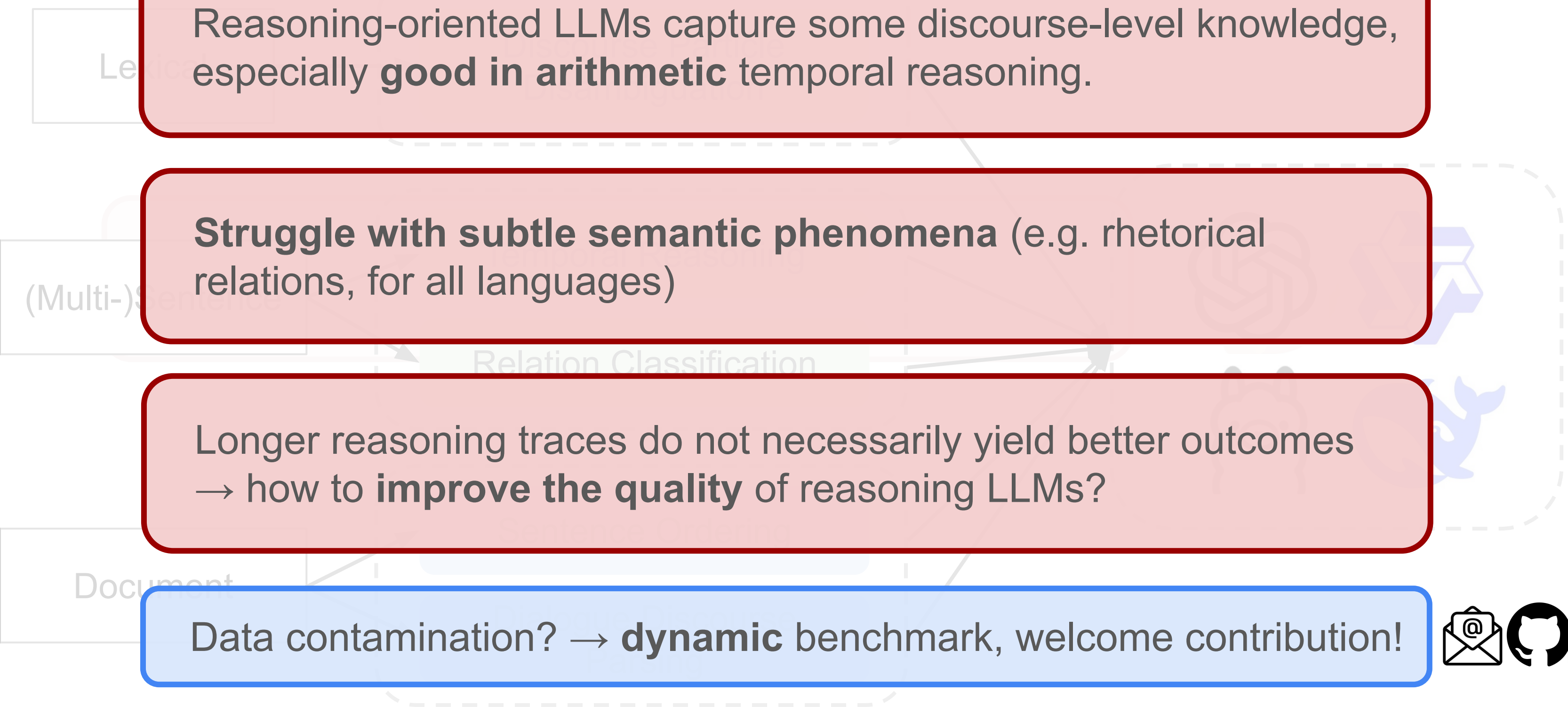
Knowledge overlap (Jaccard index)

Analysis: Inter-task correlation



- ❖ DDP – dialogue discourse parsing
- ❖ SO – sentence ordering
- ❖ TR – temporal
- ❖ DR – relation
- ❖ DM – markers

Summary: benchmark and evaluation baseline



Reasoning-oriented LLMs capture some discourse-level knowledge, especially **good in arithmetic** temporal reasoning.

Struggle with subtle semantic phenomena (e.g. rhetorical relations, for all languages)

Longer reasoning traces do not necessarily yield better outcomes → how to **improve the quality** of reasoning LLMs?

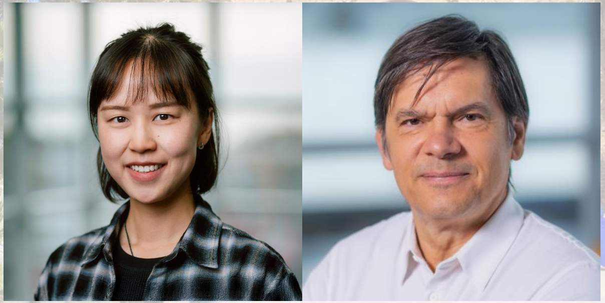
Data contamination? → **dynamic** benchmark, welcome contribution!

EACL 2026
MOROCCO
Palais Des Congres, Rabat
March 24 - 29, 2026

Thank you!

BeDiscovER: The Benchmark of Discourse Understanding in the Era of Reasoning Language Models

Chuyuan Li, Giuseppe Carenini



Thank you and welcome to our Oral Presentation at Salle Le LIXUS, Wednesday March 25 at 14:30-16:00 if you have any question or would like to chat more!