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

A broader range of discourse study with LLMs

Lexical

“recently”

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

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

“simply”

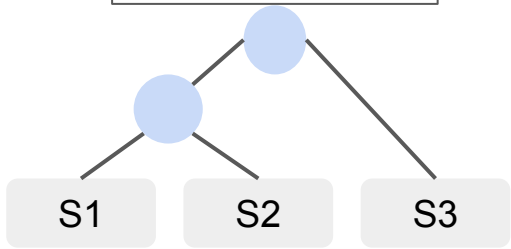
What is the sense of *just* in these contexts?

(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?

A broader range of discourse study with LLMs

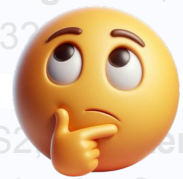
Lexical

(Multi-)Sentence

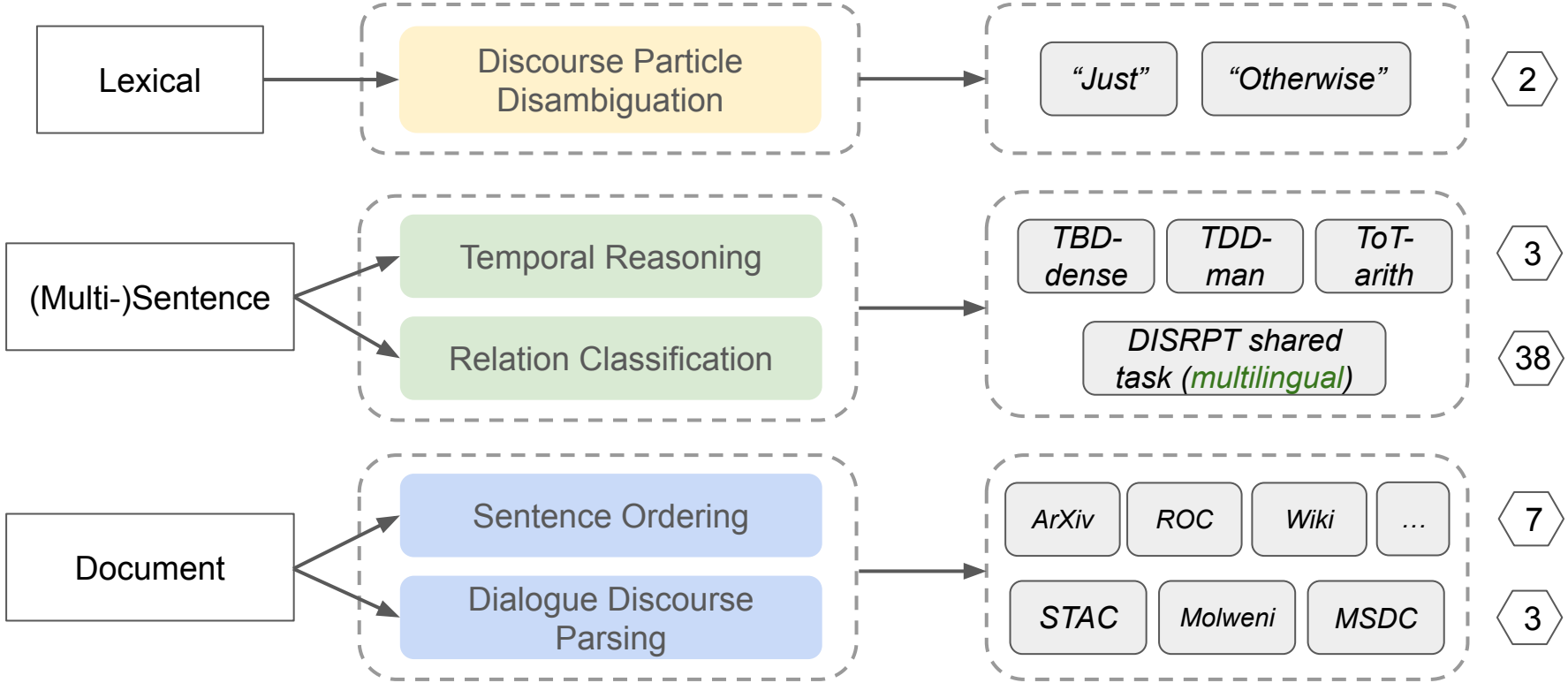
Document

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

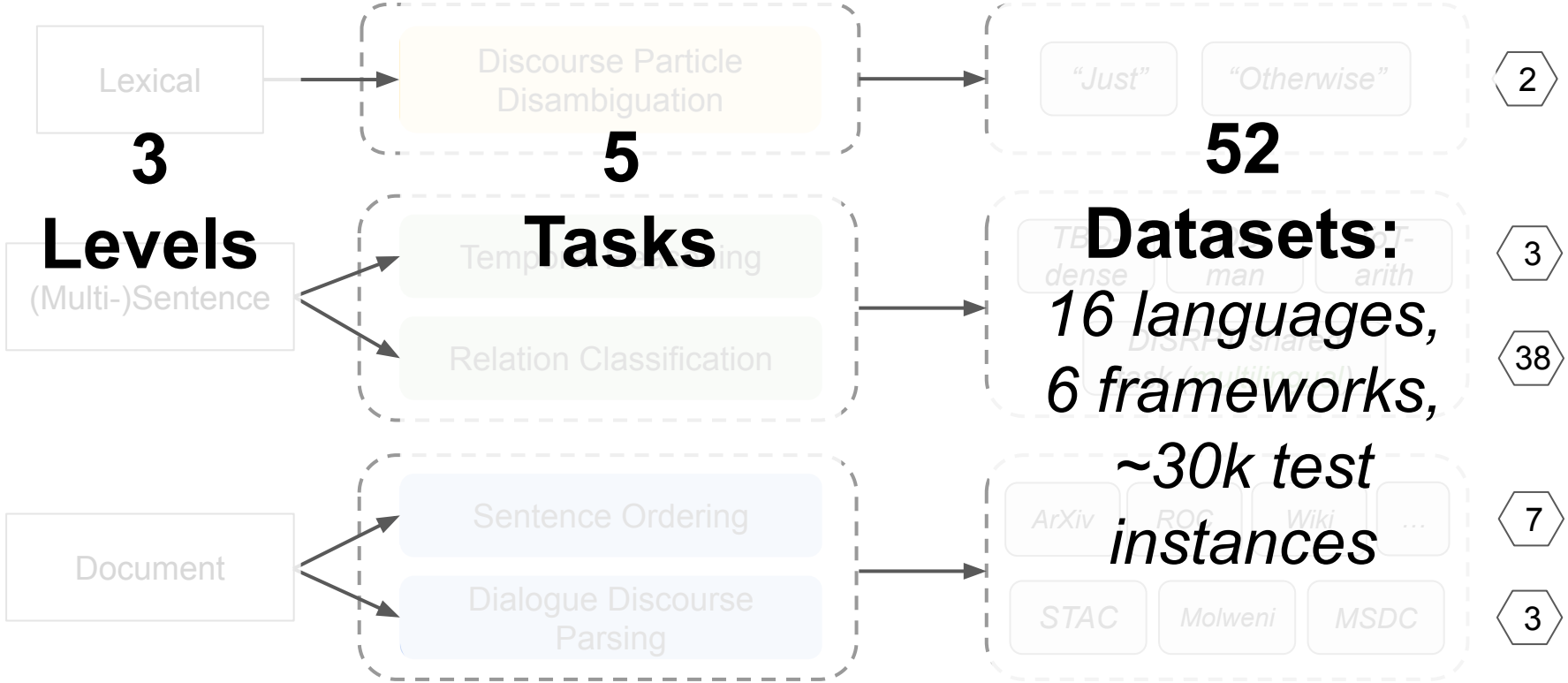
How well do modern LLMs understand discourse?



BeDiscovER: Level – Task – Dataset



BeDiscovER: Level – Task – Dataset



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

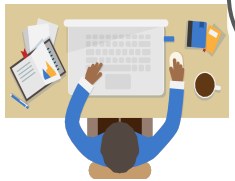
Llama-4

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?

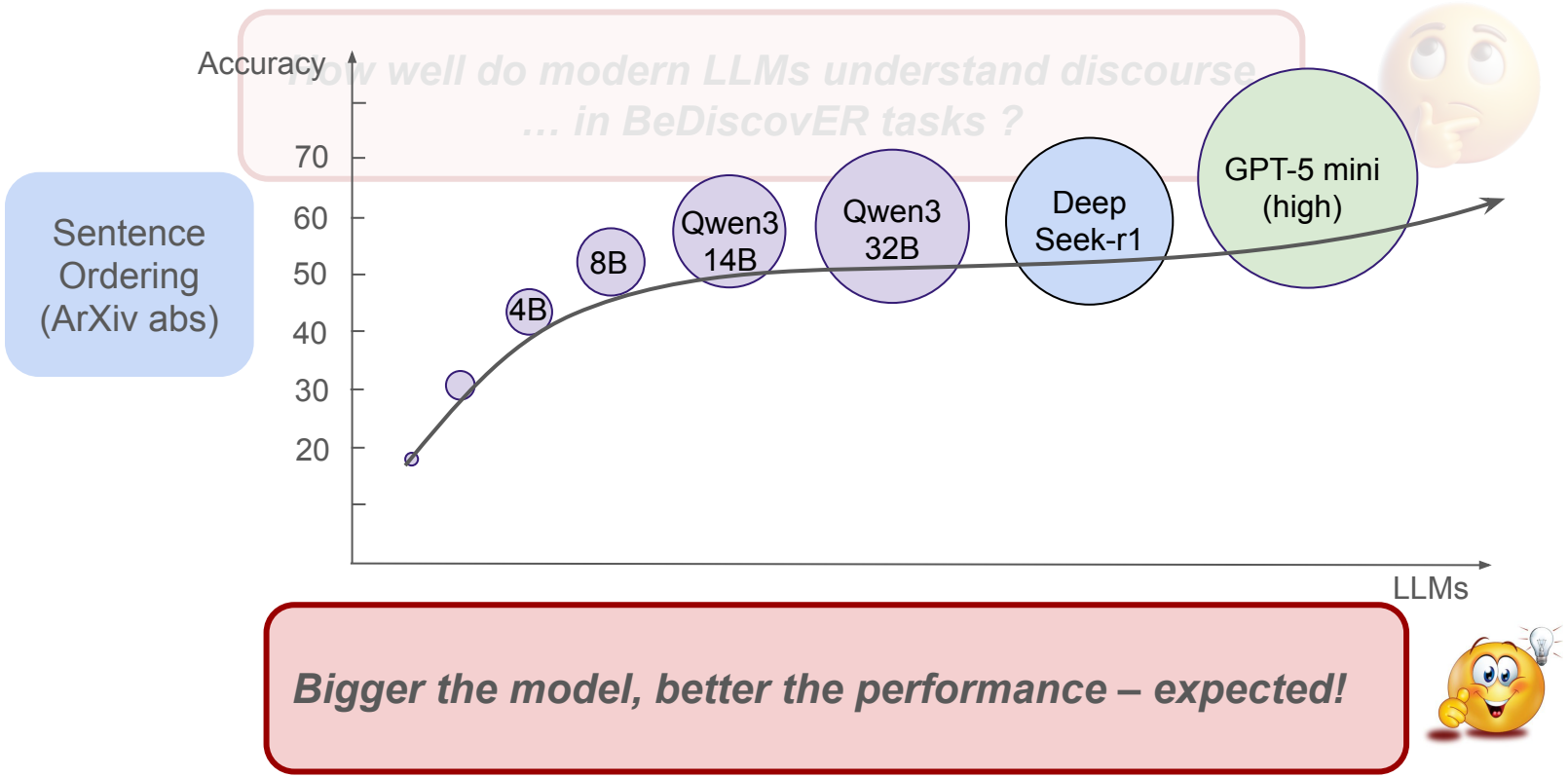


Temporal

GPT-5 Qwen3 DeepSeek-r1

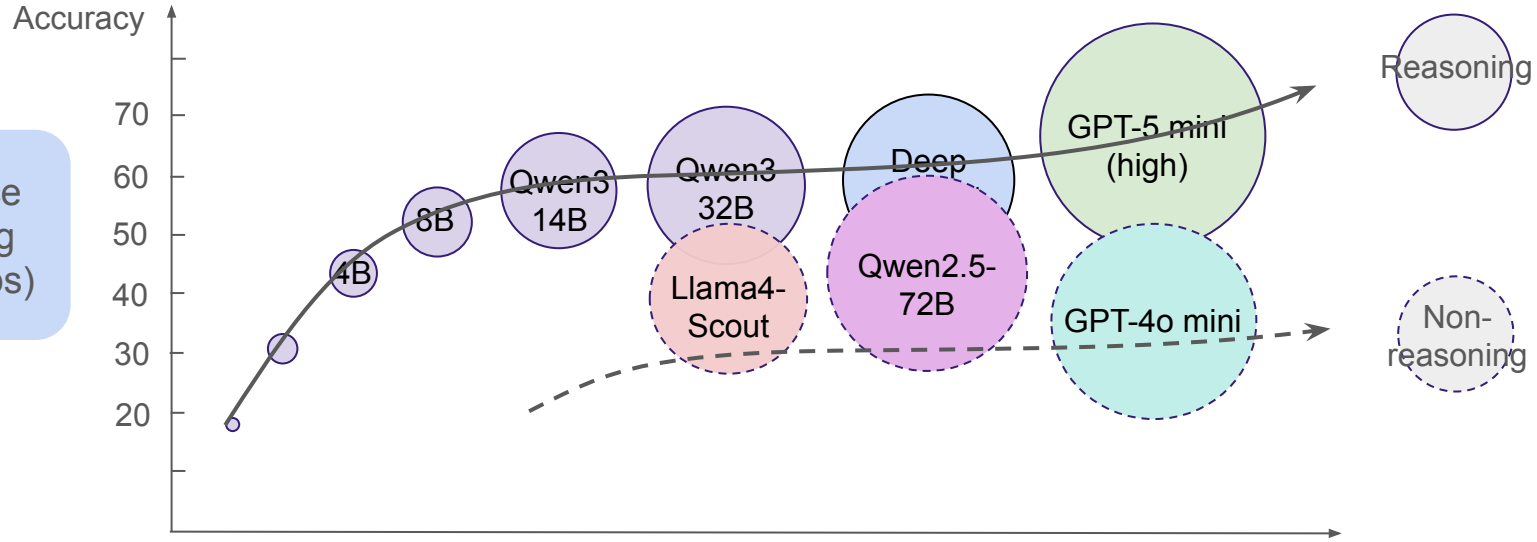
Llama-4

Performance: model scaling



Performance: reasoning-oriented vs. non-reasoning LLMs

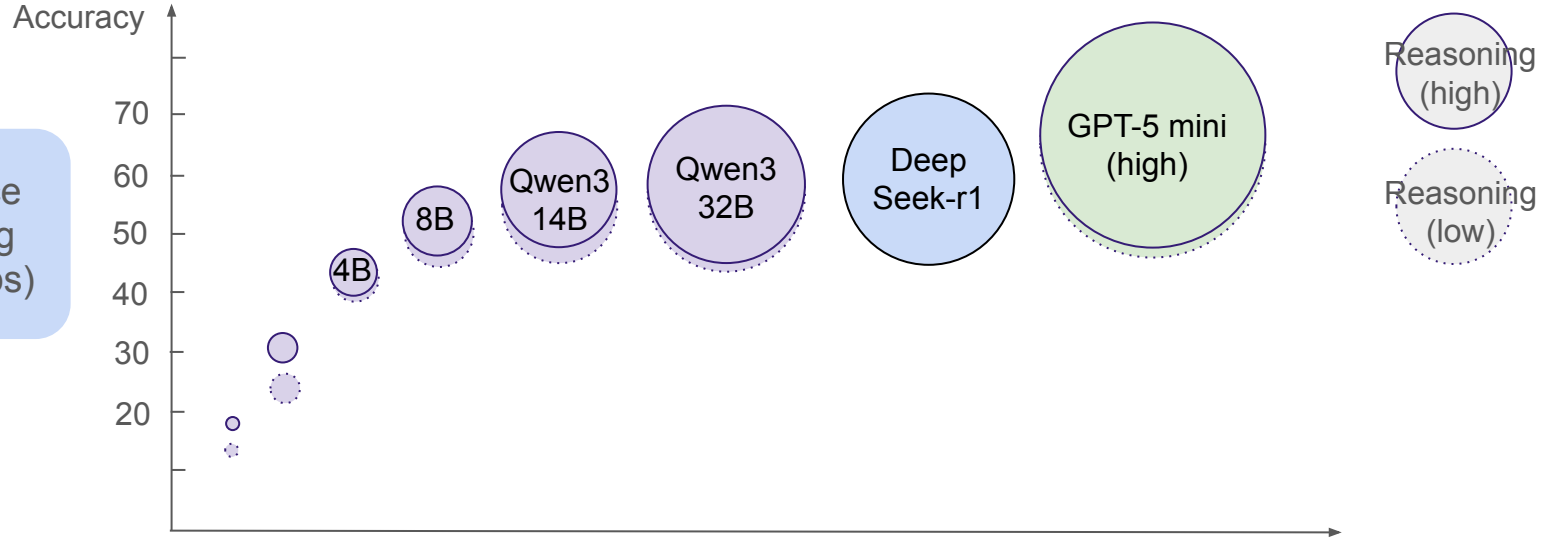
Sentence Ordering (ArXiv abs)



Reasoning-oriented LLMs outperform non-reasoning optimized LLMs.

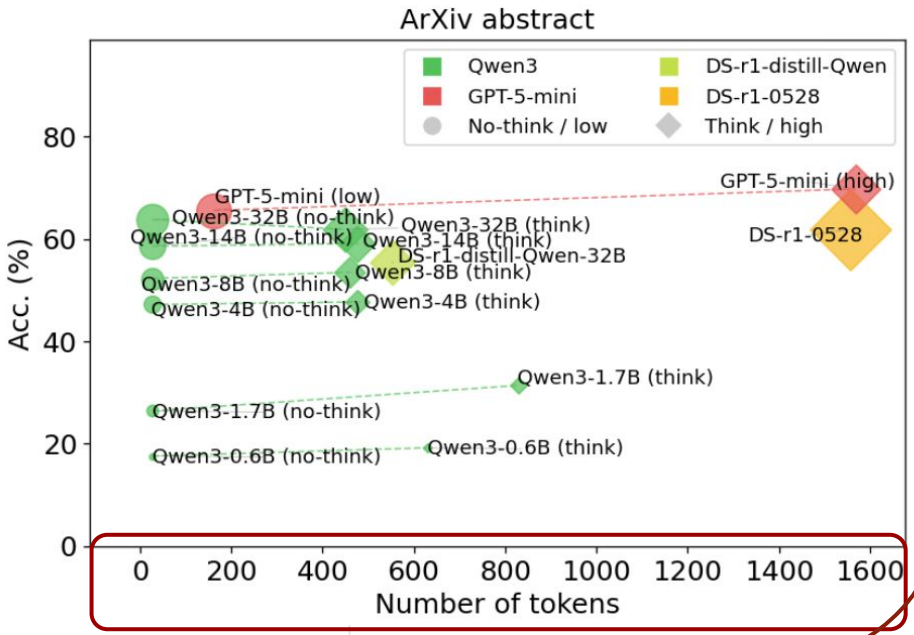
Performance: higher reasoning effort, better result?

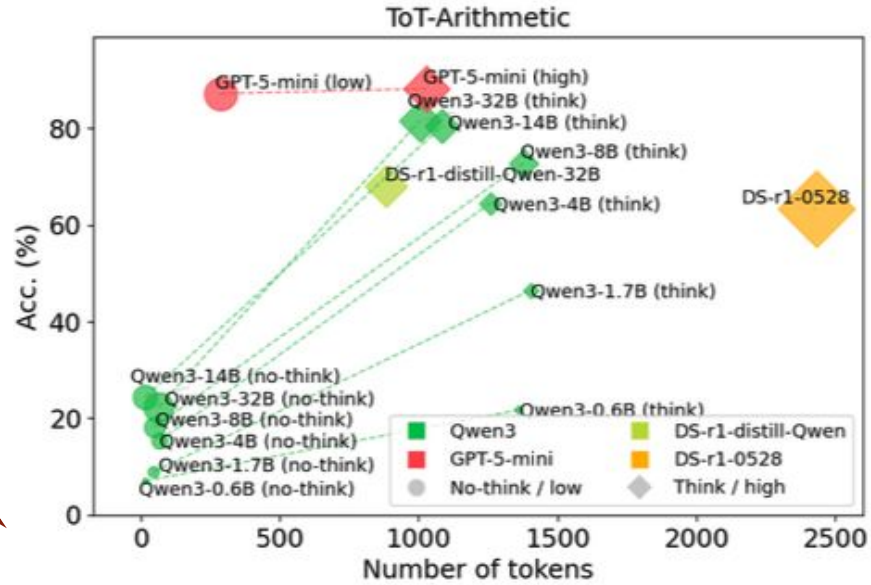
Sentence Ordering (ArXiv abs)



Higher thinking effort does not yield better outcome

Performance: higher reasoning effort, better result?

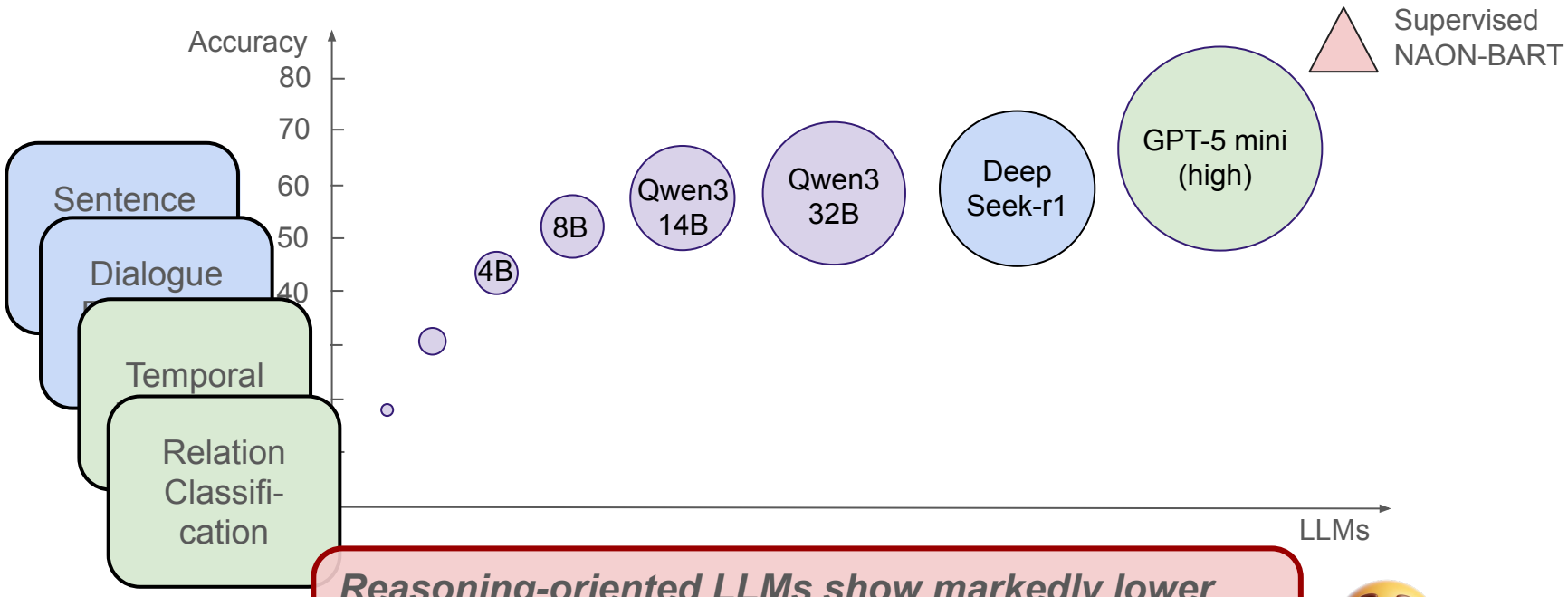




Models become more verbose without producing more meaningful reasoning…

… except for arithmetic temporal reasoning: longer reasoning lead to substantial gains!

Performance: LLMs vs. supervised

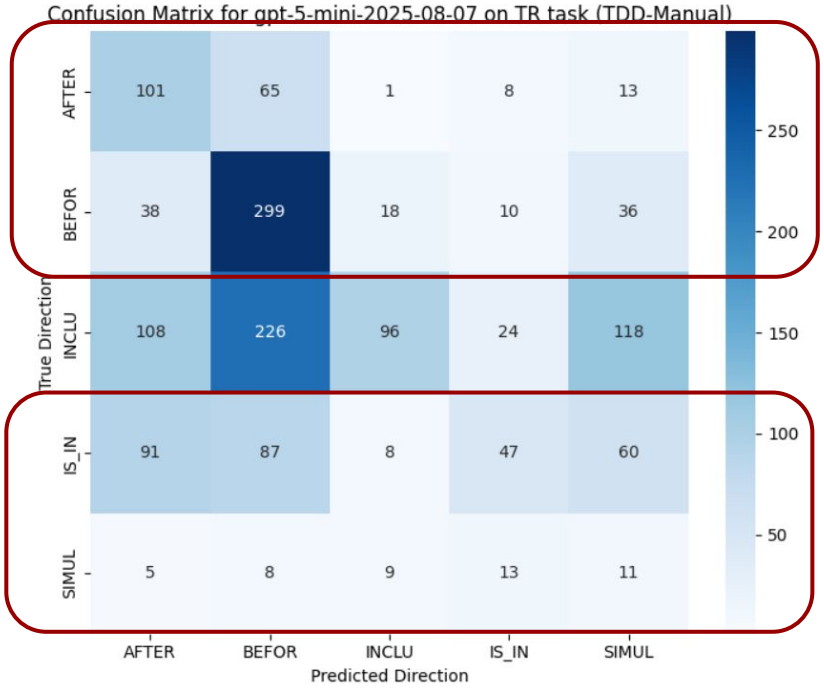


Reasoning-oriented LLMs show markedly lower performance (~10–30%) compared to supervised models, despite their larger size.

Performance: fine-grained sense disambiguation

Temporal Reasoning

Relation Classifi-cation

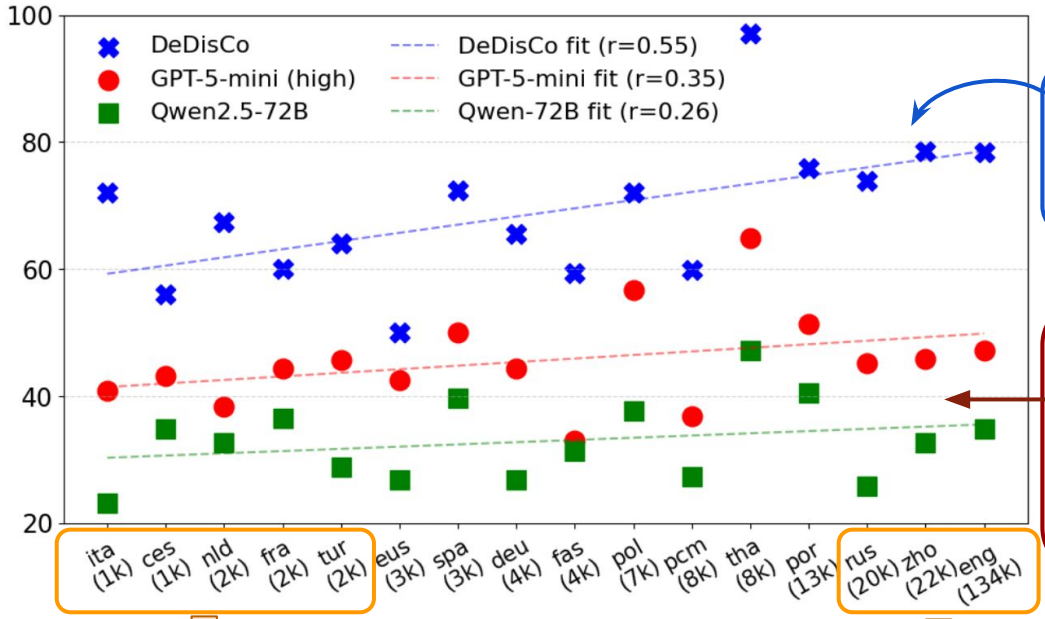


Identifies ***before/after*** relations

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

Performance: multilingual performance

Relation Classification



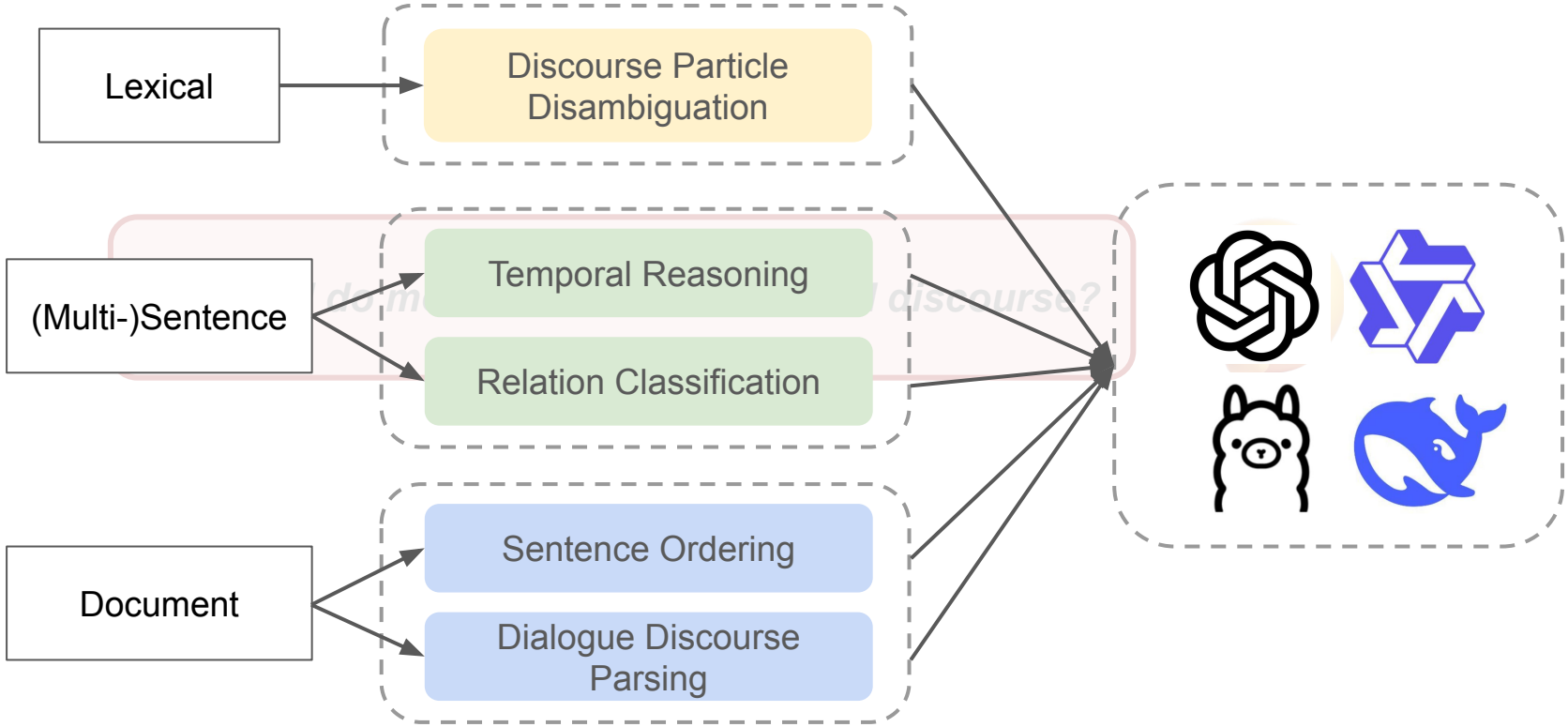
Supervised model show clear perform disparities

LLMs perform nearly flat → some generalizability & lack of robust relation representation

low-resource languages

high-resource languages

Summary: benchmark and evaluation baseline



Summary: benchmark and evaluation baseline

- Reasoning-oriented LLMs capture some discourse-level knowledge, especially **good in arithmetic aspect** of temporal reasoning.
- But they **struggle with subtle semantic and discourse phenomena** (like rhetorical relation classification) and long-dependency reasoning (dialogue parsing).
- Longer reasoning traces do not necessarily yield better outcomes in reasoning models.

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!