RESOURCEFUL 2026

LREC 20 Palma 26

From Polyester Girlfriends to Blind Mice

Creating the First Pragmatics Understanding Benchmarks for Slovene

Mojca Brglez & Špela Vintar

Faculty of Arts, University of Ljubljana
“Jožef Stefan” Institute, Ljubljana

UNIVERSITY OF LJUBLJANA | FF Faculty of Arts

Jožef Stefan Institute
Ljubljana, Slovenia

Motivation

LLMs today:

- Superior performance on more “traditional” tasks
- conversational partners

Current datasets:

- Often lack human control values
- US-/English-centric
- Machine-translated

Pragmatics: Meaning in context

implied, indirect, culturally grounded meaning

- context
- speaker intention
- shared knowledge
- cultural and social norms

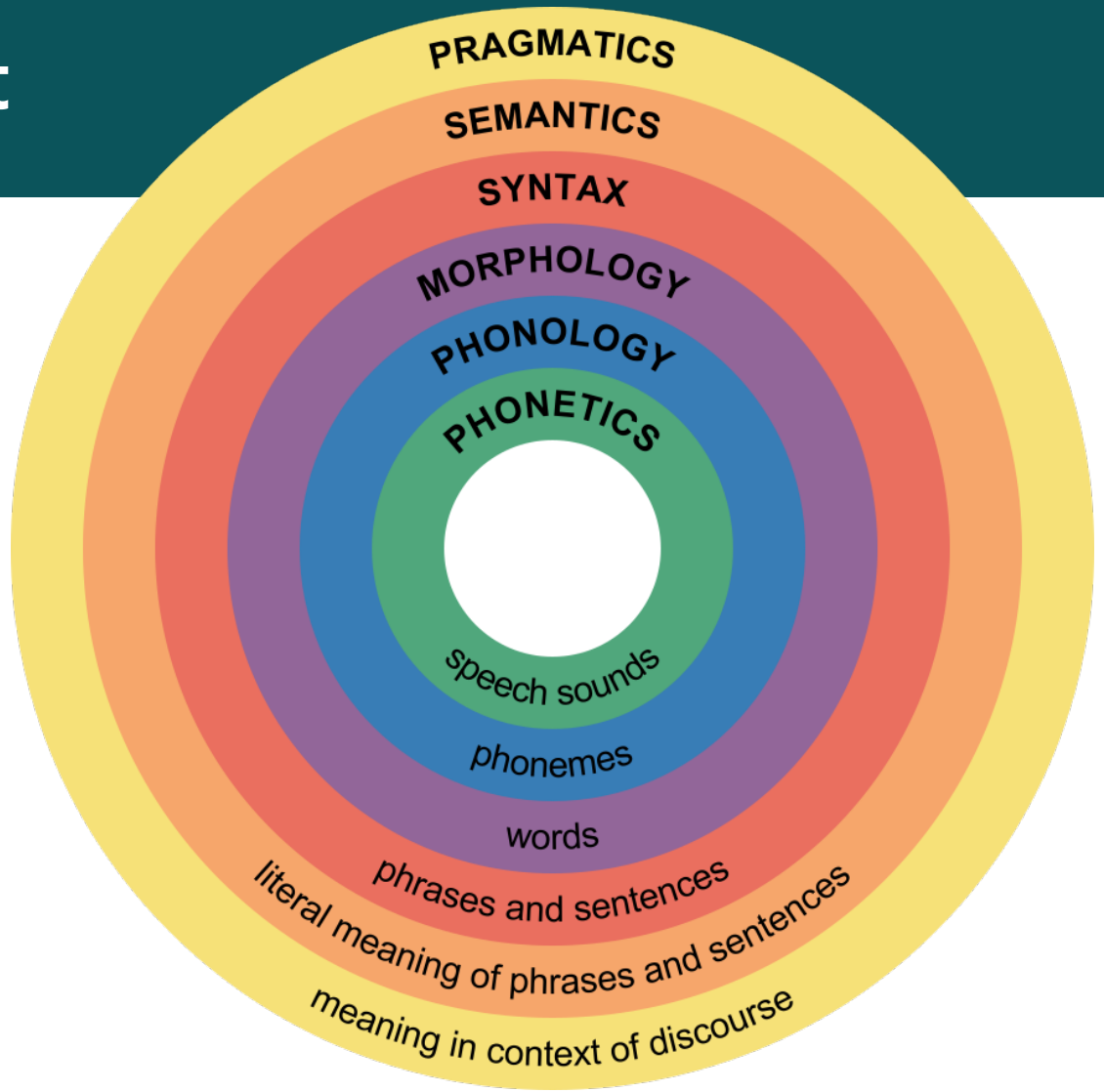


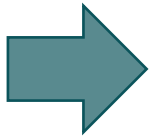
Figure: Thomas, James J. & Cook, Kristin A. , ed. (2005) *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, p. 110.]

Creating benchmarks for Slovene

Translation & adaptation of 2 multiple-choice question answering [**MCQA**] datasets

MultiPragEval (Park et al. 2024)

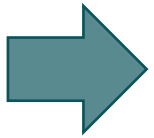
- 1st multilingual pragmatics understanding benchmark



SloPragEval

PragMega (Floyd et al. 2023)

- Multimodal behavioral test



SloPragMega

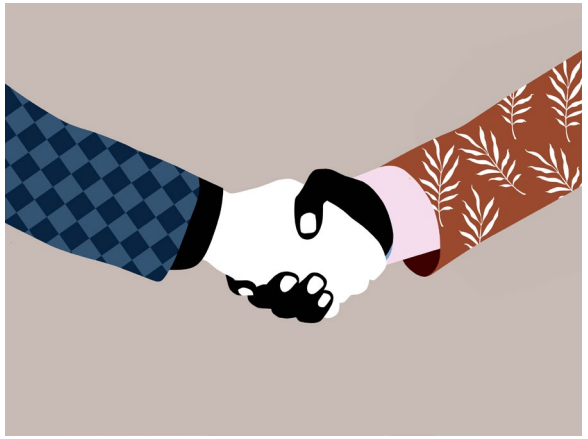
MultiPragEval

Korean + translated into English, German, Chinese:

- DeepL + human revision + *cultural adaptation

300 examples

Gricean maxim violations



The '**cooperative principle**' (Grice, 1979) & its four core maxims:

– **Quantity**: be informative,
– **Quality**: be truthful,
– **Relevance**: be relevant,
– **Manner**: be clear.

PragMega

Originally 20 tasks, 11 phenomena

Previously used in LLM evaluation (Hu et al. 2023)

Selected tasks, 100 examples total:

- **Metaphor**
- **Irony**
- **Humour**

Dataset	test	dev
SloPragEval	240	60
SloPragMega	100	5

Table 1: Benchmark dataset sizes

Localization

1. **Human translation** by translation students (MA)
2. **Revision** by translation students
3. **Revision**(s) by authors

- **Linguistic challenges** (idioms, metaphors, common phrases)
- **Culturally specific elements** (names, holidays, locations, food etc.)

PragMega Metaphor task example

Situation + Utterance

Mark asked his mom what she thought about his new girlfriend. She replied:
"This young lady is 100% polyester."

What does she mean?

5 Meaning Hypotheses

1) His girlfriend wore clothes made of polyester.
2) **His girlfriend's behavior was not very natural.**
3) The girl made a good impression on Mark's mom.
4) His girlfriend has a beautiful smile.
5) His girlfriend is made of polyester.

PragMega Metaphor task example localization

Mark asked his mom what she thought about his new girlfriend. She replied:
“This young lady is 100% polyester.”

Marko je mamo vprašal, kaj si misli o njegovem novem dekletu. Odgovorila je:
“Igra se slepe miši.”

*[lit. **‘She’s playing blind mice’**.]*

PragMega Humour
task example

Situation

A famous French mime died of a cerebral hemorrhage, the school he founded confirmed today. The doctor said:

5 Punchline
Hypotheses

A) **"He went quietly."**
B) "His talents will be greatly missed."
C) "Mime is a beautiful form of art."
D) "You can buy very good wine in France."
E) The principal of the school slipped on a banana peel and fell in front of the class.

PragMega Humour
task example

Situation

A famous French mime died of a cerebral hemorrhage, the school he founded confirmed today. The doctor said:

5 Punchline
Hypotheses

A) "He went quietly."
B) "His talents will be greatly missed."
C) "Mime is a beautiful form of art."
D) "You can buy very good wine in France."
E) The principal of the school slipped on a banana peel and fell in front of the class.

PragEval
task example

Situation + Utterance

Emily, who lives in the Middle East, returned to the United States after a long absence and spent her summer vacation at a forest resort. When her friend Charlie asked her how she liked it, Emily said:

"It was so green."

Choose the most appropriate meaning of the above utterance from the following options

5 meaning hypotheses

(A) The vacation at the forest resort was unsatisfactory.

(B) Everything was red where Emily went.

(C) Emily planted many trees with green leaves.

(D) Emily was disappointed that she went to a forest beach instead of a beach.

(E) None of the above.

PragEval
task example localization

Emily, who lives in the Middle East, returned to the United States after a long absence and spent her summer vacation at a forest resort. When her friend Charlie asked her how she liked it, Emily said:

“It was so green.”

‘in the city’ *‘a cabin in Pokljuka forest’*

Ema, ki živi v mestu, je počitnice preživela v koči sredi pokljuških gozdov. Ko jo Tadej vpraša, kako se je imela, odgovori:

“Bilo je tako zeleno.”

Human annotation campaign (SloPragEval)

Validation + baseline

- Crowdsourced via social media, total 79 participants
- Randomized samples of 50 examples

~ 10 votes per example

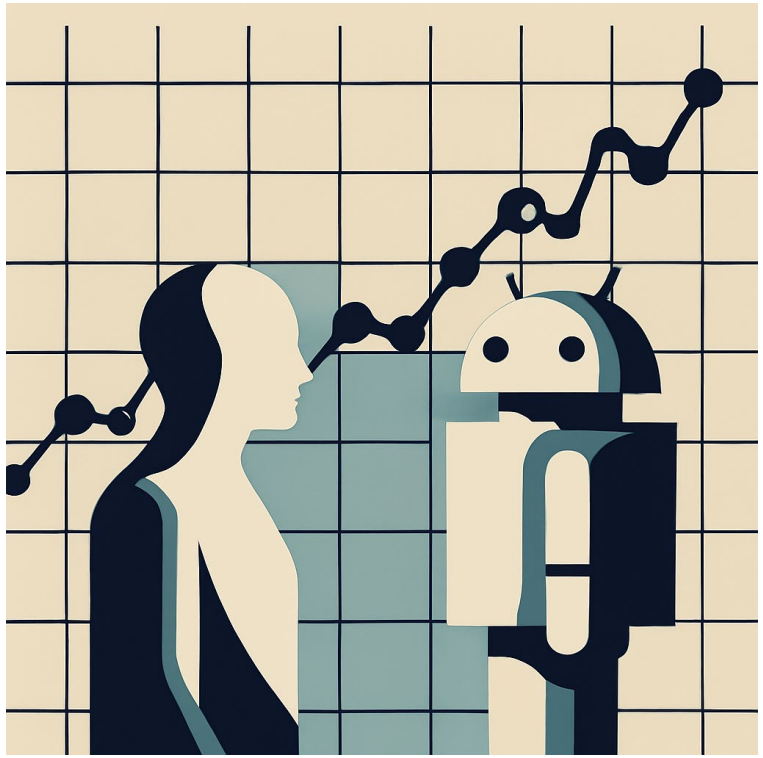


Figure generated with gpt-image-1

LLM Evaluation

- **Slovene + English** prompts: following orig. datasets; minimal instructions.

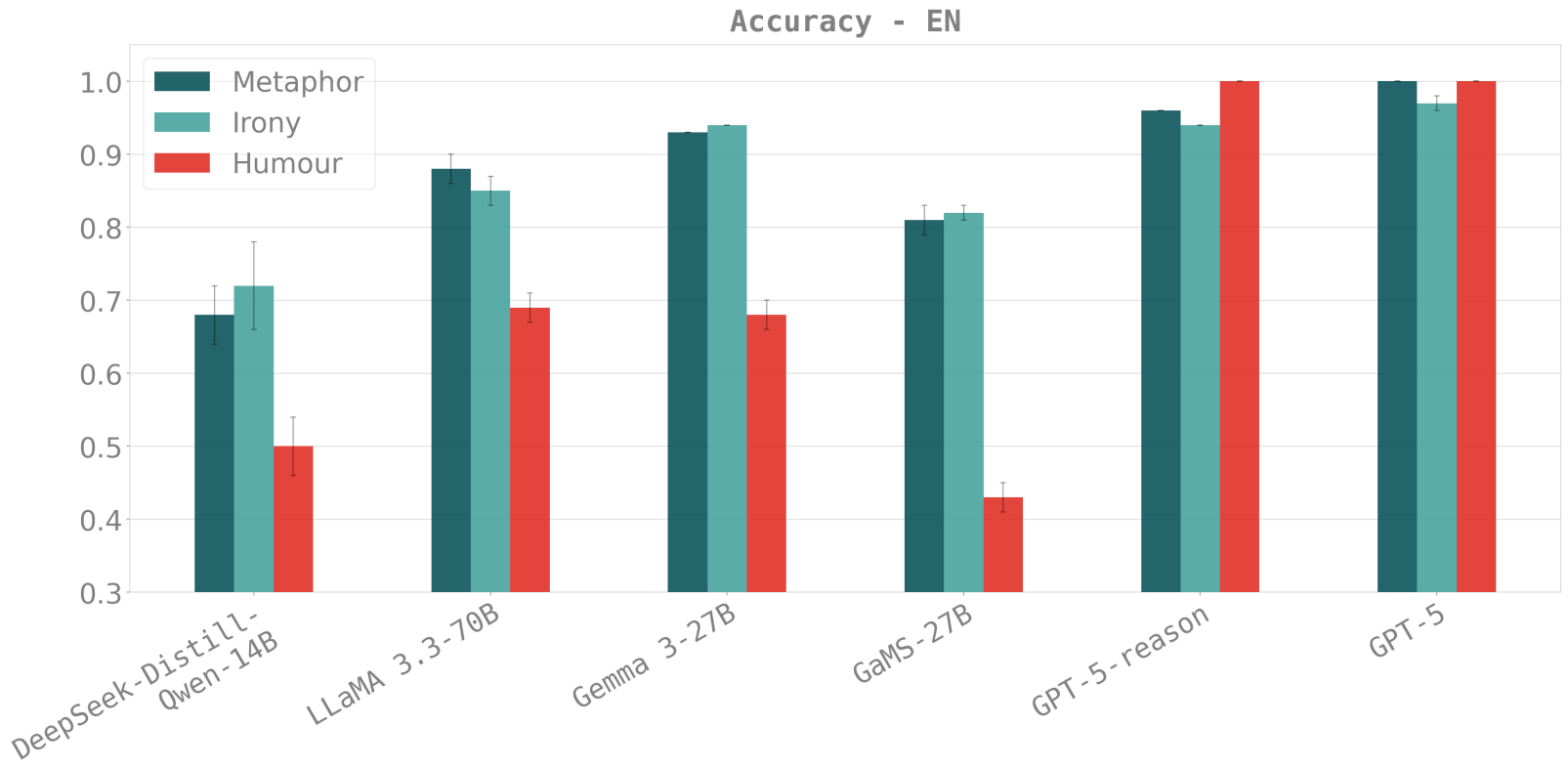
English prompt template:
Task: You will read short stories that describe everyday situations. Each story will be followed by a multiple-choice question. Read each story and choose the best answer. Your task is to decide what the character in the story is trying to convey. The answer options are 1, 2, 3, or 4.
Scenario:
[Example]
Options:
[Hypotheses]
Answer:

- DeepSeek-R1-Distill-Qwen-14B (*reasoning*)
- Gemma 3-27B
- GaMS-27B
- Llama 3.3-70B

- GPT-5
- GPT-5-chat (*reasoning*)

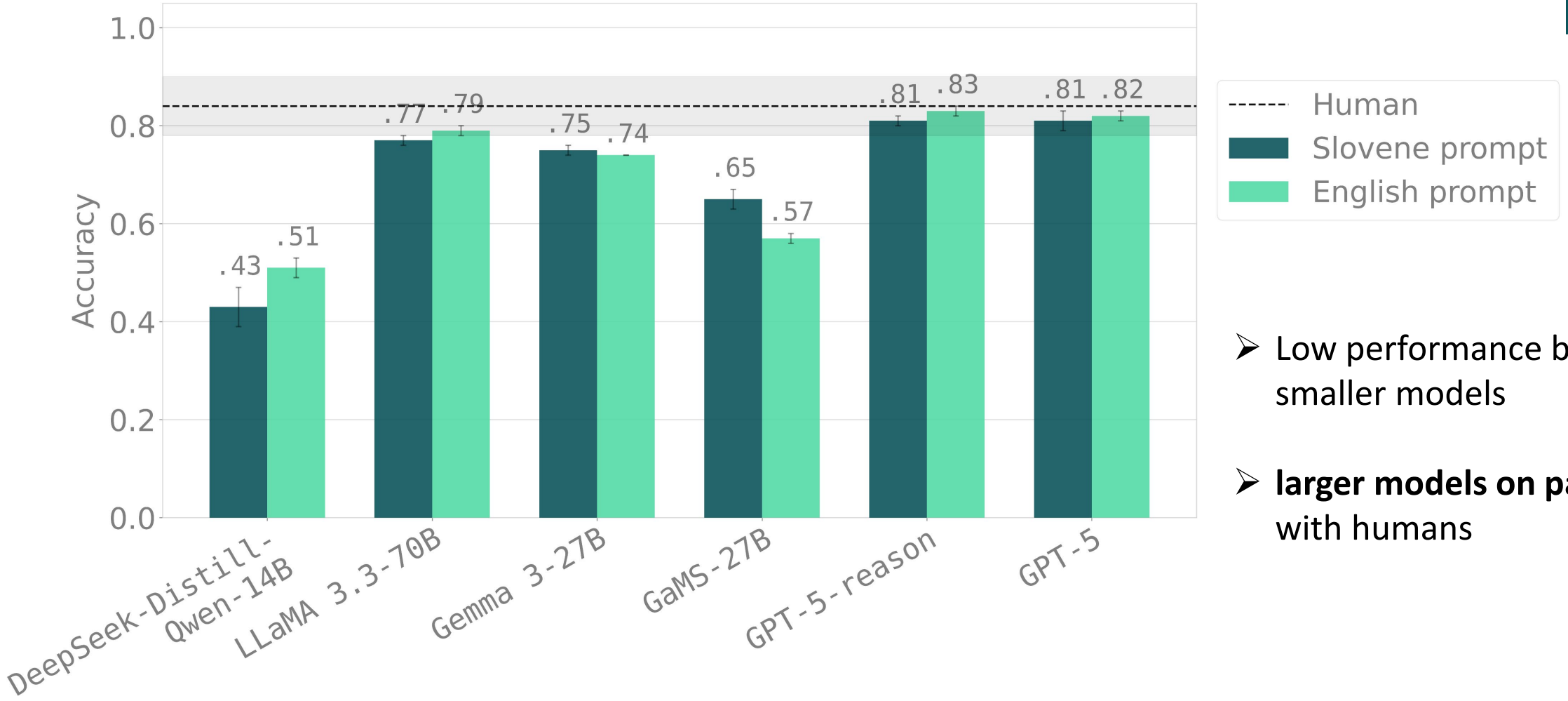
- 3 runs
- Predictions extracted by regular expressions and manually checked
- Accuracy

Results: SloPragMega



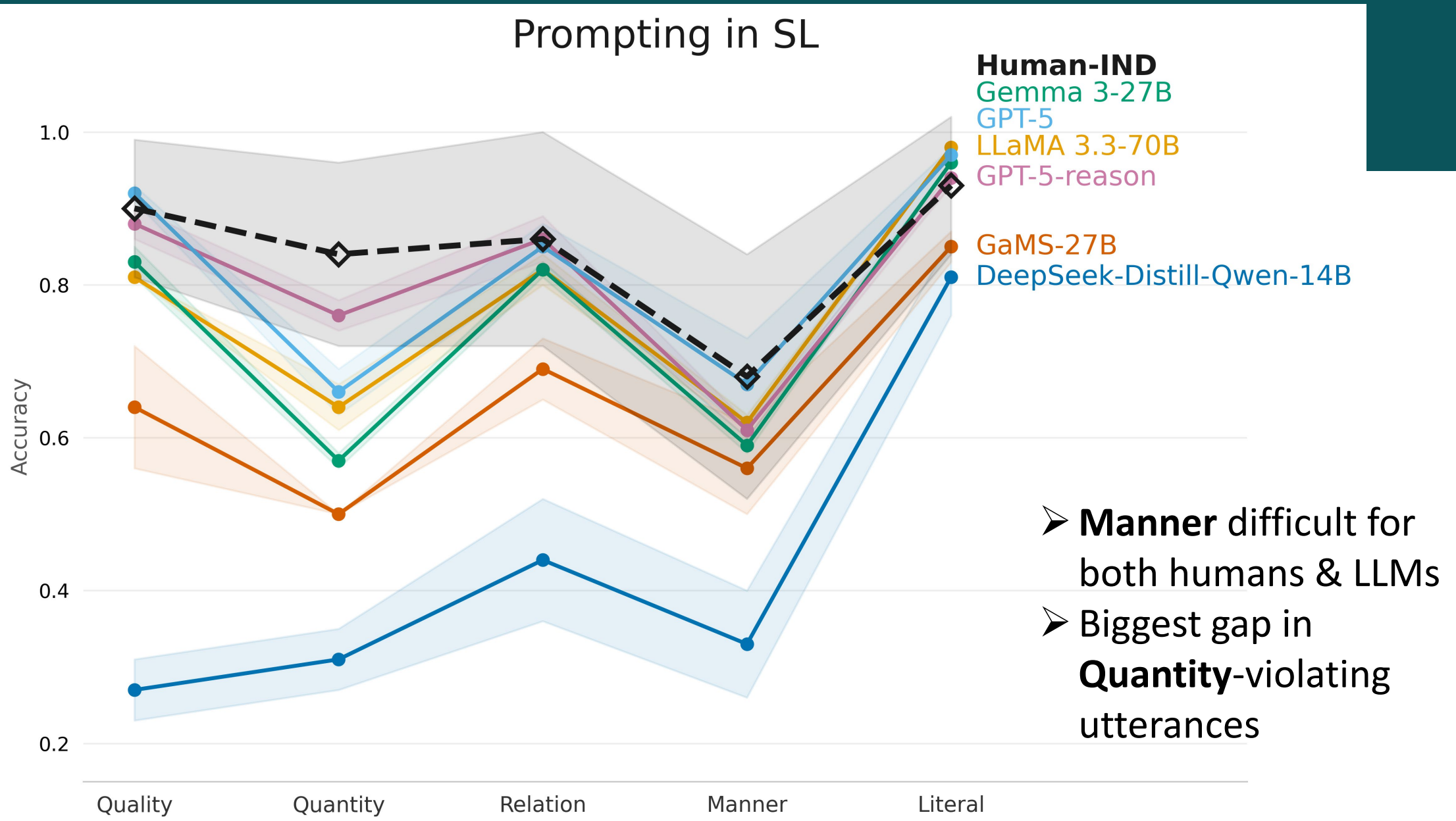
- Similar results when prompted in EN/SL
- Most challenging: **selecting humorous punchlines**
- **Significant gap** between bigger/smaller models
- **(some) perfect scores** by GPT-5!

Results: SloPragEval



- Low performance by smaller models
- **larger models on par** with humans

SloPragEval – overall average



- **Manner** difficult for both humans & LLMs
- Biggest gap in **Quantity**-violating utterances

SloPragEval – by phenomenon

Conclusions & future work

Benchmark creation

available on **SloBENCH**

difficult to translate – esp. via MT

extensive cultural and linguistic adaptation

human **validation and baseline**

- **native** Slovene data
- **dialects, code-switching, noise**
- **multimodal data**

LLM performance

closed models **approach human performance**

open models lag behind

certain phenomena remain difficult

- deeper **error analysis**
- expand **model coverage**
- **open-ended generation** of the answer

References:

- H. Paul Grice. 1975. Logic and conversation. In Cole, P., Morgan, J. (eds.) *Syntax and Semantics, vol. 3*, pp. 22–40. Academic Press, reprinted as ch. 2 of Grice 1989, 22–40.
- Dojun Park, Jiwoo Lee, Seohyun Park, Hyeyun Jeong, Youngeun Koo, Soonha Hwang, Seonwoo Park, and Sungeun Lee. 2024. MultiPragEval: Multilingual Pragmatic Evaluation of Large Language Models. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 96–119, Miami, Florida, USA. Association for Computational Linguistics.
- Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)*, pages 4194–4213. Association for Computational Linguistics.
- Floyd, Sammy and Gibson, Edward and Fedorenko, Evelina and Poliak, Moshe. 2023. *Pragmega*. OSF repository, Center for Open Science. https://osf.io/dpge6/.

This research was supported by the Slovenian research and innovation agency (ARIS) through the project ***Large Language Models for Digital Humanities****, grant number GC-0002.*