University of Colorado Boulder

Federico Ortega Riba, Annie Zaenen, Kathryn Conger, Elizabeth Spaulding, Martha Palmer

FULBRIGHT España

PUSHING ON LLMS' KNOWLEDGE OF THE CAUSED MOTION CONSTRUCTION

1. MOTIVATION

- Symbolic language analysis systems tend to have a **semi-modular approach**
- Probing LLMs to see whether they capture **real-world inferences** of these constructions
- Testing **ambiguity** in human answers and how LLMs align with these answers
- Main question: how well can LLMs interpret the mapping between the sentences that they generate and what is happening in the world?

Force-exerting verb:

"He squeezed it into the suitcase" vs.

"He squeezed it in the suitcase"

2. RELATED WORK

- Study of models' ability to **draw implications from argument structures** that are not syntactically complex but are **not canonical** with respect to their verbs
- **Caused motion construction (CMC)** introduced by Goldberg (1995). These constructions can add some sense of motion not encoded in the vocabulary
  - Causal argument
  - Argument in motion
  - Path that specifies the initial and final location
- Current systems rely on **statistical patterns** to capture lexical aspects of language, and the conventional usage outnumbers these usages that we have seen

Verb Type	Percent	Krippendorff's α	Fleiss' K
Movement	93.07%	0.8802	0.8796
Force-exerting	83.8%	0.7308	0.7295
Locative alt.	80.0%	0.6889	0.6674
Coerced	68.7%	0.5274	0.5257

Table 1. Disagreement and agreement metrics only between human subjects for movement of the direct object

3. METHODS

- 33 sentences with CMC for our experiments
- 4 types of verbs and 2 types of prepositions:
  - **Movement v.**: *She ejected him into the street*
  - **Force-exerting v.**: *She kicked him under the table*
  - **Locative alternation v.**: *He washed it from the sidewalk*
  - **Coerced v.**: *She sneezed it across the table*
  - **Directional prep.**: *into*, *through*, *toward...*
  - **Neutral prep**: *over*, *in*, *under...*
- We conducted 100 surveys to collect answers from students
- Aim: test ambiguity and reliability of linguistic judgments
- Hypotheses:
  - Linguistic students are **more likely to detect ambiguity**
  - LLMs should aim for movement or no movement

Type	Model
GPT	`gpt-4`
Gemini	`gemini-3-pro-preview & gemini-3-flash-preview`
Llama4	`Maverick-17B-128E-Instruct-FP8 & Scout-17B-16E-Instruct-FP8`

Table 2. LLMs used for the task

4. RESULTS

- Locative alternation and coerced verbs are **ambiguous** for LING students
- **Neutral prepositions** are more difficult to interpret
- LLMs tend to **disambiguate better**, especially for coerced verbs
- Movement verbs: **very high agreement** for both humans and LLMs' answers. Only *eject under* was problematic
- Force-exerting verbs: *kick*, *push* and *squeeze* are ambiguous, usually with **neutral prepositions**
- Locative alternation verbs: *feed*, *stuff* and *hose* are **ambiguous for both LLMs and humans**
- Coerced verbs: *laugh*, *jeer*, *glare* and *snore* have **poor human agreement** regardless verb or preposition

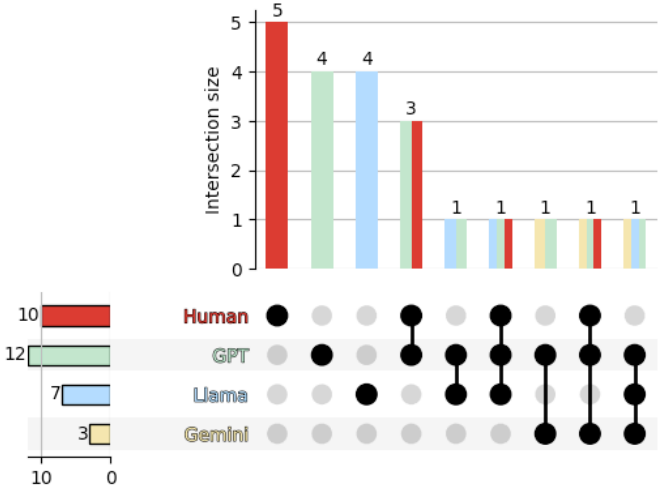

Fig 1. Intersection of wrong or ambiguous interpretations between LLMs and human judgements

REFERENCES