Do language models have coherent mental models of everyday things?

By: Yuling Gu and Bhavana Dalvi Mishra and Peter Clark

NATURAL LANGUAGE PROCESSING · MASTER'S DEGREE IN PHYSICS OF DATA

Gloria Isotton - 2072705
Kostas Panagiotakis - 2081260

May 2024

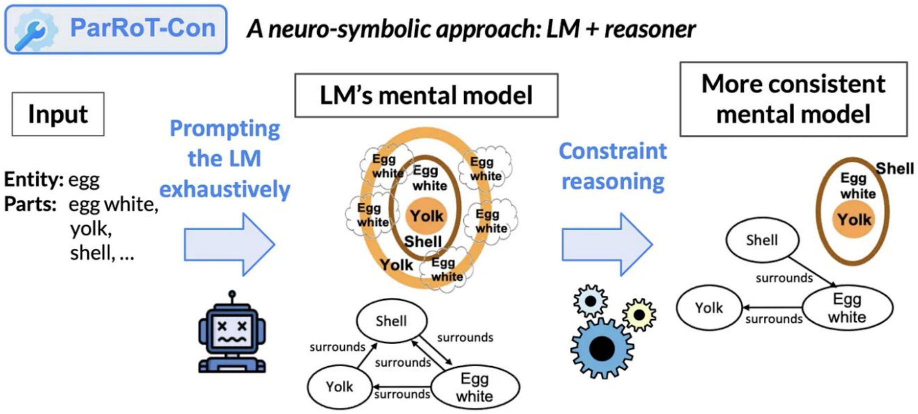
UNIVERSITÀ DEGLI STUDI DI PADOVA

Introduction

- People effortlessly understand the parts and relationships of everyday objects. The study aims to **assess** if language models (**LMs**) possess similar **coherent understandings**.

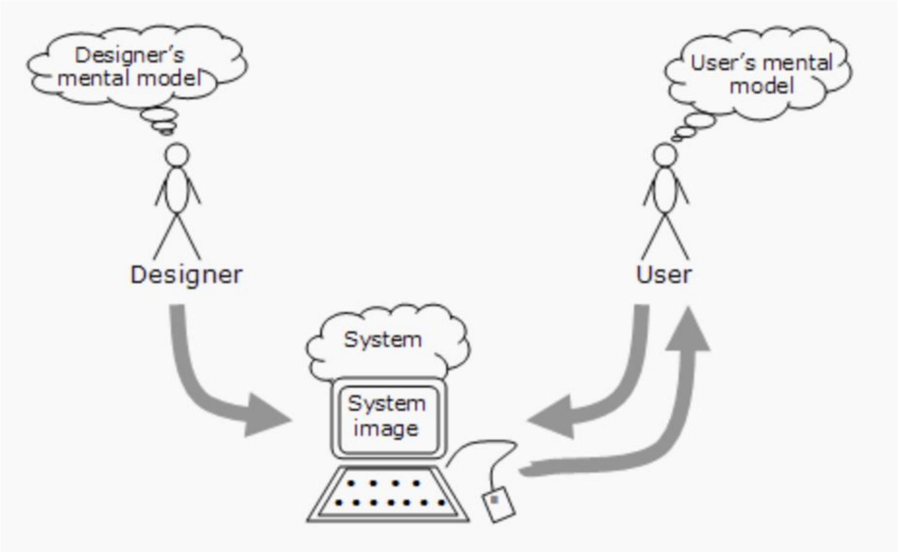
- **Dataset** Description: Consists of 100 **everyday objects**. Includes **parts** and **relationships** expressed as 11,720 true/false questions.

- Proposed Solution: Integrate a **constraint satisfaction layer** onto LM's raw predictions to enhance coherence while applying commonsense constraints.



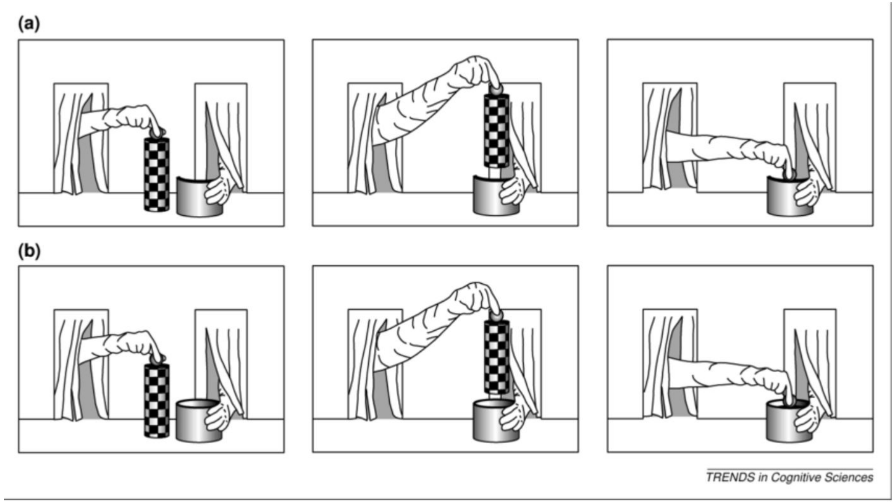
Related Work

- Craik described mental models as "**small-scale models**" of external reality
- Johnson-Laird emphasized the importance of **coherent internal representations** of spatial layouts in human reasoning



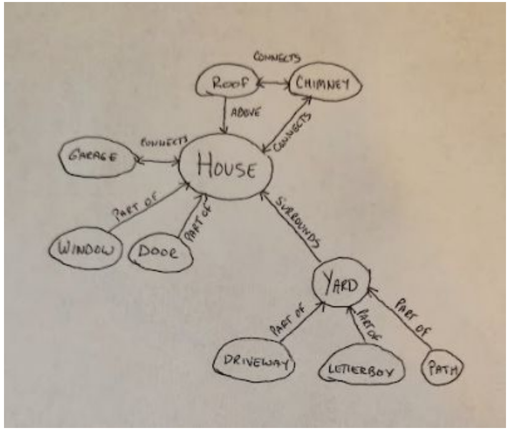
Do LMs possess coherent internal representations?

- Psychologists suggest:
  - **humans** develop **mental models** of the **world** to base decisions on (Ha and Schmidhuber, 2018; Jonassen and Henning, 1996).
  - **humans** develop **mental models** of the **world** to base decisions on exhibit understanding of object properties before language comprehension (Hespos and Spelke, 2004).
- State-of-the-art LMs like **GPT-3** and **Macaw perform poorly** in answering relationship queries between object parts.



Building and Evaluating Parts Mental Models

- "**Parts mental model**", (pmm), for everyday objects: a focused subset of a complete mental model represented as a **directed graph**. Parts mental models consist of **nodes** representing **parts** and **edges** indicating **relationships** between them, including:
  - **spatial orientation**
  - **connectivity**
  - **functional dependency**
- Dataset creation involves **human annotators** constructing **pmm** based on predefined parts and relationship vocabulary.
- The dataset as an ensemble is known as **ParRoT**.
- Considered **14 different relationships** between parts of objects



https://www.dropbox.com/sh/tv2hc6pmsbr25l3/AAAXZKvfkfyx6SAkgjolhS0ra?dl=0&e=1&file_subpath=%2FParRoT_MM_sketches&preview=ParRoT_MM_sketches.zip

Type	Relations
Spatial orientation	part of, has part, inside, contains, in front of, behind, above, below, surrounds, surrounded by, next to*
Connectivity	directly connected to*
Functional dependency	requires[2], required by

A closer look: the Task

- Define the objective as **constructing** a **parts mental model** for **everyday things**.
- **Input** Specifications: Include the everyday thing, parts list, and a relation vocabulary comprising 14 relations.
- **Output** Specifications: Require a list of tuples (x, r, y) where relation r holds between parts x and y.
- **Evaluation Criteria**: Specify the criteria for evaluating LM-generated parts mental models, including **accuracy** of LM-generated parts mental models compared to gold-standard models in our dataset and adherence to **commonsense constraints** such as the inverse relation between 'above' and 'below', ensuring predictions align with established patterns.

Enriched annotations

- The model in the study automatically infers **new connections** (positive & negative) between everyday objects, using four types of constraints (symmetric, asymmetric, inverse, and transitive)
- Inferred relationships can be **positive** (e.g., "A above B" implies "B below A") or **negative** (e.g., "A above B" implies "B NOT above A").
- This adds over **11,700 relationships** to the knowledge base.

	Given as seed (unique)	Annotated mental models	Avg. annotated per mental model	Annotated + enriched (*) (Total)	Total avg. per mental model (Total / # mental models)
# everyday things	100	100	-	100	-
# mental models	-	300	-	300	-
# parts	716	2191	7.30	2191	7.30
# relations (p1, rln, p2)	8	2752	9.17	11720	39.07
# spatial relations	6	1858	6.19	9956	33.19
# connectivity relation(s)	1	818	2.73	1612	5.37
# functional relation(s)	1	76	0.25	152	0.51

ParRoT-Con: Improving LM

Step 1: Probing a Pre-trained Language Model

Query:

Judge whether this statement is true or false:
In a tree, roots are surrounded by the trunk.

part 1 relation part 2

LM
GPT 3/ Macaw :

Answer:

{True; False}

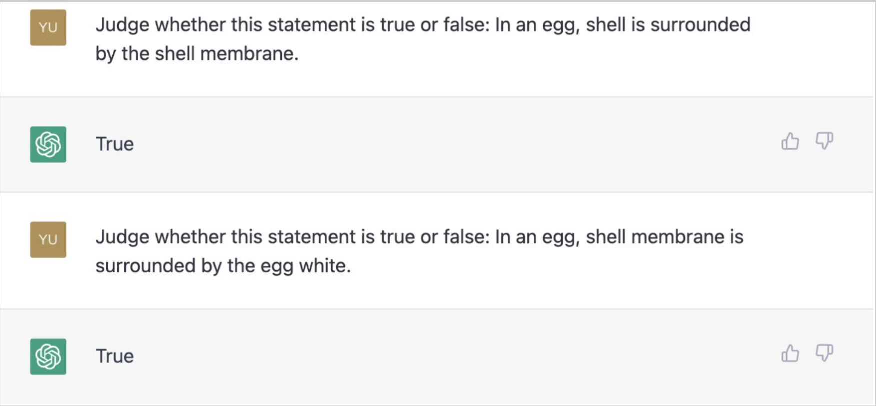
ParRoT-Con: Improving LM

Step 2: **Constraint Reasoning**

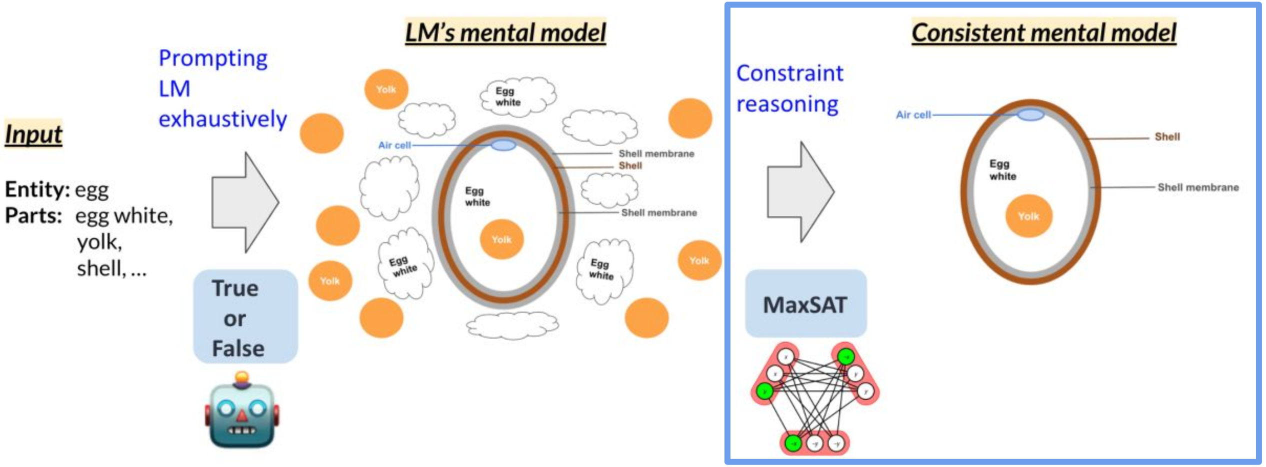
Many *inconsistencies* have been reported by considering the following hard constraints:

Constraint	**Math Symbols**
Symmetric relations	x rln $y \leftrightarrow$ y rln x
Asymmetric relations	x rln $y \lor$ y rln x
Inverse relations	x rln $y \leftrightarrow$ y inverse(rln) x
Transitive relations	x rln $y \land$ y rln $z \rightarrow$ x rln z

ChatGPT: **violation of Asymmetric relation**



Step 2: Constraint Reasoning



- **Constraint reasoning refines predictions** from the LM using weighted **MaxSAT Solver;**
- it tries different combinations of true/false values for the relation tuples to fulfill:
  - **Soft Clauses** (preferences): preserve the model's raw answers;
  - **Hard Clauses** (mandatory): minimize constraint violations;

What is accuracy?

True/False accuracy compared to the 11.7K gold relation tuples present in ParRoT

GPT-3	Judge whether this statement is true or false: In a tree, roots are surrounded by the trunk.	True (incorrect)
GPT-3	Judge whether this statement is true or false: In a tree, trunk is below the roots.	False (correct)

- baseline accuracy at 59% ;
- random chance at 50%.

Base LM results

	# params	Base LM (%)
GPT-3 (text-davinci-003)	175B	53.83
Macaw-11B	11B	59.45

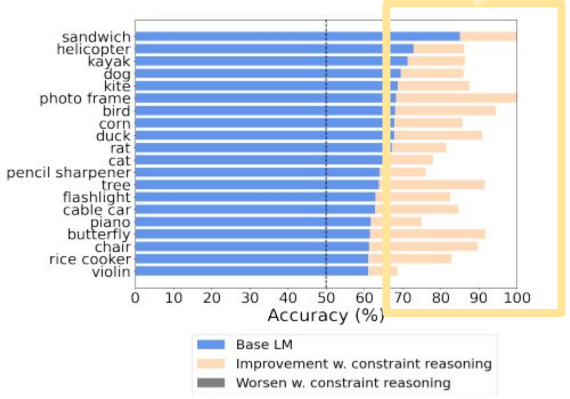
AVERAGE ACCURACY

ParRoT-Con results

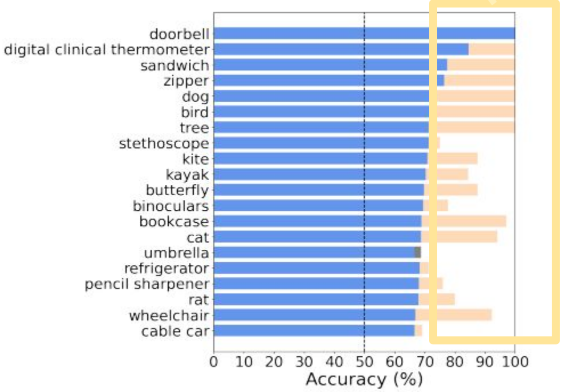
	# params	ParRoT-Con (%)	Improve (%)
GPT-3 (text-davinci-003)	175B	70.26	16.42
Macaw-11B	11B	79.28	19.84

AVERAGE ACCURACY

20 everyday things that each model achieved best performance on.



(a) GPT-3



(b) Macaw-11B

What is (in)consistency?

Measure inconsistency across the 4 types of constraints:

$$\tau = \frac{\sum_{x \in D} \left[\bigvee_{(L,R)} \neg(L(x) \rightarrow R(x)) \right]}{\sum_{x \in D} \left[\bigvee_{(L,R)} L(x) \right]}$$

$$= \frac{\text{size of the set of violated constraints}}{\text{size of the set of applicable constraints}}$$

Base LM results

- 19-43% **conditional violation** on average;

AVG CONDITIONAL VIOLATION

	%True tuples	%Conditional Violation (lower is better) Symmetric relations	Asymmetric relations	Inverse relations	Transitive relations	Avg. (macro)	Avg. (micro)
GPT-3 (text-davinci -003)	12.64	66.37	23.01	71.14	32.18	48.17	42.84 (27,105/63,265)
Macaw-11B	57.77	29.98	64.97	33.63	10.08	34.66	19.23 (111,022/577,322)

- GPT-3 struggles with symmetric and inverse relations consistency;
- Macaw-11B struggles with asymmetric relations;

ParRoT-Con results

Produces perfectly consistent mental models for all LMs with respect to the imposed constraints i.e. **0 % conditional violation** for all columns in table.

Conclusions

The authors were able to:

1. **Built a Benchmark**: dataset, ParRoT. This dataset includes detailed information about 100 common things, outlining over 2,000 parts and the relationships between them (more than 11,700 relationships in total).
2. **Exposed LM Weaknesses**: Using ParRoT, the authors showed that current LMs generally lack strong mental models of everyday objects, violating basic common sense constraints of everyday things;
3. **Introduced ParRoT-Con**: develop a method, ParRoT-Con, to solve the inconsistency problem, which has proven to improve both accuracy (up to 20% improvement) and consistency.