SemEval 2024 BRAINTEASER: A Novel Task Defying Common Sense

Task Home Page

Motivation

Human reasoning processes comprise two types of thinking: vertical and lateral. Vertical thinking, also known as linear, convergent, or logical thinking, is a sequential analytical process that is based on rationality, logic, and rules. Meanwhile, lateral thinking (or "thinking outside the box") is a divergent and creative process that involves looking at a problem from a new perspective and defying preconceptions.

The success of language models has inspired the NLP community to attend to tasks that require implicit and complex reasoning, relying on human-like commonsense mechanisms. While such vertical thinking tasks have been relatively popular, lateral thinking puzzles have received little attention. To bridge this gap, we devise BRAINTEASER: a multiple-choice Question Answering task designed to test the model's ability to exhibit lateral thinking and defy default commonsense associations.

BRAINTEASER QA task consists of two subtasks-**Sentence Puzzle** and **Word Puzzle** that require awareness of commonsense "defaults" and overwriting them through unconventional thinking that distinguishes these defaults from hard constraints.

- **Sentence Puzzle**: Sentence-type brain teaser where the puzzle defying commonsense is centered on sentence snippets.
- **Word Puzzle**: Word-type brain teaser where the answer violates the default meaning of the word and focuses on the letter composition of the target question

Both tasks include an adversarial subset, created by manually modifying the original brain teasers without changing their latent reasoning path.

Task Example

Here are two examples from each subtasks

Question	Choice
A man shaves everyday, yet keeps his beard long.	**He is a barber.**
	He wants to maintain his appearance.
	He wants his girlfriend to buy him a razor.
	None of the above.
What part of London is in France?	**The letter N.**
	The letter O.
	The letter L.
	None of the above.

To ensure that our task evaluates reasoning ability rather than memorization, we construct adversarial versions of the original data in two ways:

- **Semantic Reconstruction** rephrases the original question without changing the correct answer and the distractors.
- **Context Reconstruction** keeps the original reasoning path but changes both the question and the answer to describe a new situational context.

Here are the example of two adversarial versions of Sentence Puzzle:

Adversarial Strategy	Question	Choice
Oringinal	A man shaves everyday, yet keeps his beard long.	**He is a barber.** He wants to maintain his appearance. He wants his girlfriend to buy him a razor. None of the above.
Semantic Reconstruction	A man preserves a lengthy beard despite shaving every day.	**He is a barber.** He wants to maintain his appearance. He wants his girlfriend to buy him a razor. None of the above.
Context Reconstruction	Tom attends class every day but doesn't do any homework.	**He is a teacher.** He is a lazy person. His teacher will not let him fail. None of the above.

Each system will be evaluated based on the following two accuracy metrics:

- **Instance-based Accuracy**: We considers each question (original/adversarial) as a separate instance. We will report accuracy for the original, and its adversarials.
- **Group-based Accuracy**: Each question and its associated adversarial instances form a group, and a system will only receive a score of 1 when it correctly solves all questions in the group.

Data

The pilot data is now available. The training and validation split will be releaded based on the SemEval timeline.

Registration form for participation and the legal usage of data.

Mailing list for task updates.

For further question, please contact: yifjia@isi.edu

Codalab

The tasks are set to be facilitated on CodaLab, with the availability of the link being aligned with the SemEval schedule. Participants are encouraged to register at the earliest and join the mailing list to stay abreast of updates.

Leaderboard

Sentence Puzzle

Team	Original	Semantic	Context	Ori & Sem	Ori & Sem & Con	Overall

Team						
ChatGPT (zero-shot)	60.77	59.33	67.94	50.72	39.71	62.68

Word Puzzle

Team	Original	Semantic	Context	Ori & Sem	Ori & Sem & Con	Overall
ChatGPT (zero-shot)	56.10	52.44	51.83	43.90	29.27	53.46

Important Dates

Event	Date
Tasks announced (with sample data available)	17 July 2023
Training data ready	4 September 2023
Evaluation start	10 January 2024
Evaluation end by by	31 January 2024 (latest date; task organizers may choose an earlier date)
Paper submission due	29 February 2024
Notification to authors	1 April 2024
Camera ready due	22 April 2024
SemEval workshop	TBD, 2024 (co-located with a major NLP conference)

Organization

Name	Affiliation	Email
Yifan Jiang	USC,ISI	yifjia@isi.edu
Filip Ilievski	USC,ISI	ilievski@isi.edu
Kaixin Ma	CMU, LTI	kaixinm@andrew.cmu.edu