MultiZebraLogic: A Multilingual Logical Reasoning Benchmark

Sofie H. Bruun, Dan Saattrup Smart
The Alexandra Institute
Rued Langgaards Vej 7, 2300 København
{sofie.bruun,dan.smart}@alexandra.dk

Dataset on Hugging Face in 9 languages

Code for data generation and evaluation

Summary

- **Multilingual logical reasoning benchmark** designed for both reasoning and non-reasoning LLMs, covering **9 Germanic languages**.
- Source code for **automatic "zebra puzzle" generation** built for scalability to more languages or themes. Zebra puzzles require multi-step reasoning to solve. They can be solved analytically, but this is slow for humans. No LLM's are used during puzzle generation.
- 2x3 and 4x5 puzzle sizes are suitably difficult for GPT-4o mini and o3-mini, respectively. These models represent **a non-reasoning and a reasoning model**.
- **Red herrings** (non-informative clues) increase difficulty.
- No clear difference in difficulty between clue types, **English vs. Danish**, or between the **classic theme vs. a culture-specific theme.**
- To be implemented in **EuroEval leaderboards.**

Puzzle Generation

1. Generate a random solution matrix.

	Attributes (in categories)			
Objects	object_1	police officer	fantasy	handball
	object_2	nurse	romance	bouldering

2. Choose clues.

while not solved:

Suggest a random clue – use an applicable clue type and insert objects and attributes from the solution matrix

Check the number of possible solutions N_i

If $N_i < N_{i-1}$:

Keep the clue

If $N_i = 1$:

Try removing each clue

solved=True

3. Choose red herrings with random, irrelevant information.

4. Complete the puzzle text based on language and theme. Input grammar and phrases are reviewed by native speakers. E.g. the attribute "baker" is described by "the baker", "is a baker" or "is not a baker" in English.

Linguistic priorities

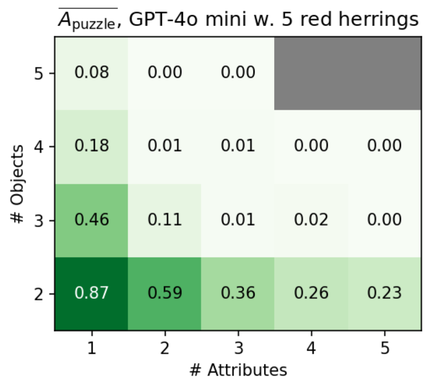
1. Correctness. Text must be linguistically acceptable.
2. Unambiguity. Clues must represent a unique solution.
3. Naturalness. Phrases should sound typical of the chosen language.
4. Ease of generation. Puzzle generation should be simple.
5. Consistency. Text should be consistent in meaning and form across languages.
6. Diversity. A variety of properties and clue types should be included.

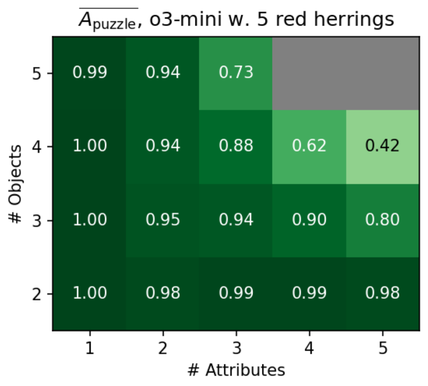
Results

- Code and dataset of 2x3 and 4x5 puzzles. LLM's are prompted to output the solution matrix given a puzzle and format instructions.
- The reasoning model o3-mini performs much better than GPT-4o mini on large puzzles. GPT-4o mini solves 36±5 % of 2x3 puzzles, and o3-mini solves 42±5 % of 4x5 puzzles.

Evaluation metrics

- Puzzle-level accuracy, A_{puzzle}: 1 if solved, 0 otherwise
- Cell-wise accuracy, A_{cell}: Fraction of correctly guessed cells in the solution matrix





- Adding 5 red herrings decreases mean A_{puzzle} by 15±7 percentage points for 4x5 puzzles with o3-mini.
- No clear correlation of scores and included types of clues or red herrings.
- No clear difference in scores for English vs. Danish or the classic houses theme vs. a culture-specific smørrebrød (open sandwiches) theme.

		Danish smørrebrød	Danish houses	English houses
A_{puzzle}	Mean	0.42±0.05	0.33±0.05	0.40±0.05
	Sample standard deviation	0.5	0.5	0.5
A_{cell}	Mean	0.66±0.04	0.66±0.04	0.67±0.04
	Sample standard deviation	0.4	0.4	0.4

Clue Types

Clue type	Positional Constraint	Example
found_at	$X = P$	The person who plays board games lives in house no. 2.
not_at	$X \neq P$	The science fiction reader does not live in house no. 1.
same_object	$X = Y$	The police officer reads crime novels.
not_same_object	$X \neq Y$	The dog owner does not like apples.
next_to	$\lvert X - Y \rvert = 1$	The zebra owner lives next to the person who loves strawberries.
not_next_to	$\lvert X - Y \rvert > 1$	The person who boulders does not live next to the person who loves blackcurrants, and they are different people.
just_left_of	$Y - X = 1$	The teacher lives to the immediate left of the rabbit owner.
just_right_of	$X - Y = 1$	The teacher lives to the immediate right of the coffee drinker.
left_of	$X < Y$	The rabbit owner lives to the left of the person who plays board games.
right_of	$X > Y$	The Brit lives to the right of the romance reader.
between	$X < Y < Z \vee X > Y > Z$	The person who loves blackcurrants lives between the police officer and the person who loves wild strawberries.
not_between	$\neg(X < Y < Z \vee X > Y > Z) \wedge X \neq Y \wedge X \neq Z \wedge Y \neq Z$	The rabbit owner does not live between the coffee drinker and the juice drinker, and they are three different people.
one_between	$\lvert X - Y \rvert = 2$	There is one house between the Norwegian and the police officer.
multiple_between	$\lvert X - Y \rvert = N_{between} + 1$	There are 2 houses between the nurse and the baker.

Data Example

A row of houses have numbers 1 to 2 from left to right.

In each house lives a person with unique attributes in each of the following categories:

Jobs: nurse and police officer.
Favourite book genres: fantasy and romance.
Hobbies: bouldering and handball.

We also know the following:

1. The police officer lives to the left of the nurse.
2. The person who plays handball does not live in house no. 2.
3. The romance reader lives in house no. 2.
4. The person with glasses does not live in house no. 1.

Who has which attributes and lives in which house?

A simple 2x3 puzzle with one red herring.

TrustLLM
ALEXANDRA INSTITUTTET
Danish Foundation Models
Funded by the European Union