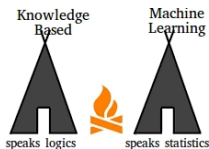


Solving logical puzzles with ChatGPT

Adrian Groza
Department of Computer Science,
Technical University of Cluj-Napoca, Romania



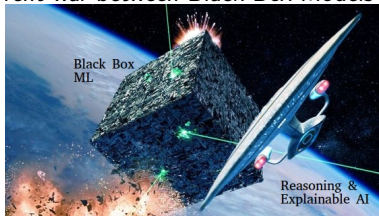
Machine Learning Alliance

The Romulan Star Empire: Supervised Learning
The Klingon Empire: Unsupervised Learning
The Borg Collective: Reinforcement Learning

All these successful villains speak the same language: **Statistics**

“Resistance Is Futile” (F-measure 0.99999)

The current war between Black Box Models and XAI



Reasoning speed:

Slow (knowledge driven)

Fast (data driven)



Who is this?

Who was his predecessor?

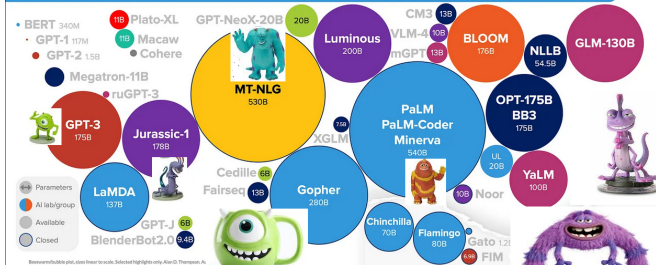
Who was his predecessor's predecessor?

Computer science also differs from physics in that it is not actually a science. It does not study natural objects. Neither is it, as you might think, mathematics; although it does use mathematical reasoning pretty extensively. Rather, computer science is like engineering

Richard Feynman

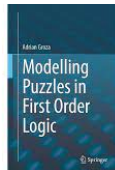


LANGUAGE MODEL SIZES TO AUG/2022



- 1 Can LLMs (ChatGPT, BARD) solve logic puzzles?
- 2 Which are the types of logical faults? How many?
- 3 Can LLMs translate into FOL?

144 puzzles (12 puzzles × 12 chapters)



Chapter 1 - Micro arithmetic puzzles

Puzzle 3. Logic equation 5×5

In this 5×5 logic equation you have to find unique integer values for the variables A, B, C, D, E - ranging from 1 to 5 - to make all statements true: (©Brainzilla - www.brainzilla.com)

$$C = A + E$$

$$E = B + 2$$

$$(B * E + 3 * E) \neq B \rightarrow A * A + D > E$$

Chapter 2 - Strange numbers

Puzzle 14. Multiplication

How many solutions are for: $A B C D E F * 3 = B C D E F A$, where each digit is distinct? (puzzle from Math is fun - www.mathisfun.com [Pierce \(2020\)](#))

Chapter 3 - Practical puzzles

Puzzle 30. Golomb ruler

Define a ruler with $M = 4$ marks (e.g. a, b, c, d) so that the distances between any two marks are different. Your ruler should be able to measure all the integer distances up to length $L = 6$. There should be only one way of measuring an integer distance with your ruler.



Puzzle 45. Ninth day: three rooms

One room contains a lady and the other two contain tigers. At most one of the three signs is true. The sign on the first room says: "A tiger is in this room". The sign on the second room says: "A lady is in this room". The sign on the third room says: "A tiger is in room 2". Which door to open in order to find the lady? **Smullyan (2009)**



Chapter 5 - Einstein or zebra puzzles

Puzzle 55. Perfect man

Susan's perfect man has black hair, brown eyes, and is tall and slim. Susan knows 4 men: Arthur, Bill, Charles and Dave. Only one of them has all the characteristics that Susan requires.

1. Arthur and Bill have the same eye colour.
2. Only one of the men has both black hair and brown eyes.
3. Bill and Charles have the same hair colour.
4. Only two of the men are both tall and slim.
5. Charles and Dave are of differing build.
6. Only two of the men are both tall and dark-haired.
7. Dave and Arthur are the same height.
8. Only three of the men are both slim and brown-eyed.

Who is Susan's perfect man? (taken from **Clessa (1996)**)

Chapter 6 - Island of truth

Puzzle 61. We are both knaves

On the island of knights and knaves, knights always tell the truth, while knaves always lie. You are approached by two people. The first one says: “We are both knaves”. What are they actually? (Smullyan (2011))



Chapter 7 - Love and marriage

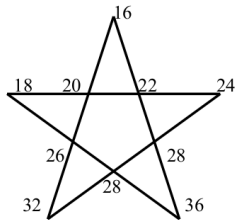
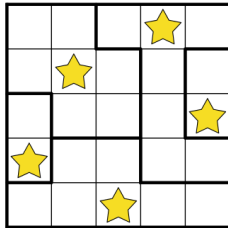
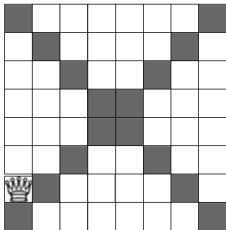
Puzzle 81. Two single persons at the end of the row

Four married men and three unmarried men are seated in a row at random. What are the chances that the two men at the ends of the row will be single? (adapted from puzzle 470 from Dudeney (2016))

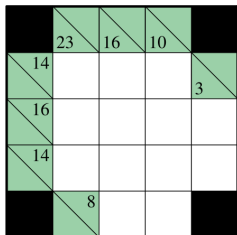
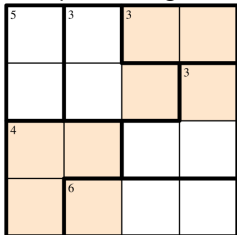


Puzzle 95. Fancy queens

I have placed a queen in one of the white squares of the board shown. Place 7 more queens in white squares so that no 2 of the 8 queens are in line horizontally, vertically, or diagonally (adapted from puzzle 113 from Kordemsky 1992).



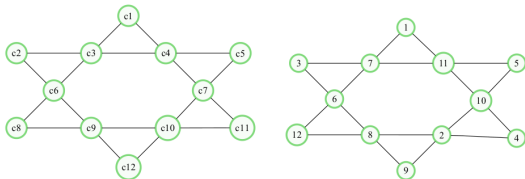
Chapter 9 - Japanese puzzles, e.g. Killer Sudoku, Kakuro



Chapter 10 - Russian puzzles

Puzzle 119. A star

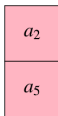
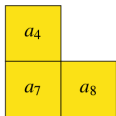
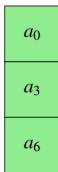
Can you place the integers from 1 through 12 in the circles of the six-pointed star so that the sum of the numbers in each of the six rows is 26? (puzzle 324 from Kordemsky (1992))



Chapter 11 - Polyomino puzzles

Puzzle 122. A simple polyomino

This puzzle uses one monomino, one domino, and two trominoes, for a total of nine squares. Assume that you cannot rotate the shapes. Group the four shapes in a 3×3 grid.



a_0	a_1	a_2
a_3	a_4	a_5
a_6	a_7	a_8

Chapter 12 - Self reference puzzles

Puzzle 133. Tricky messages

Which of the following statements are true?

1. The next two statements are false.
2. The last statement is false.
3. The next statement is false.
4. The first statement was false. (puzzle from Walicki (2016))

Puzzle 100. Self-counting sentence

You have to insert numbers in the blanks to make the following sentence true:

In this sentence, the number of occurrences of 0 is __, of 1 is __, of 2 is __, of 3 is __, of 4 is __, of 5 is __, of 6 is __, of 7 is __, of 8 is __, and of 9 is __.

There are exactly two solutions. (puzzle created by Douglas Hofstadter (Hofstadter (1985)))

Dataset of experiments

100 puzzles, ChatGPT, BARD

Solution by ChatGPT

Wrong

To solve the puzzle, we need to find two arrangements of the five cards (3, 1, 5, 7, 9) in a row such that the number formed by the first pair multiplied by the number formed by the last pair, with the central number subtracted, will produce a number composed of repetitions of one figure. Task understanding: correct Let's consider the possible pairs for the first and last numbers. Solving strategy: Analysing all possibilities

- (3, 9): The product of 39 and 17 is 663, but subtracting 5 would give us 658, which does not have repetitions of a single figure.
- (1, 7): The product of 17 and 59 is 1003, ~~but subtracting 5~~ Lack of commonsense: 5 is no longer the central number, but 3 would give us 998, which also does not have repetitions of a single figure.
- (5, 9): The product of 59 and 37 is 2183, ~~and subtracting 7~~ Lack of commonsense: 7 is no longer the central number, but 1 gives us 2176, ~~which is composed of repetitions of the figure 6~~ Wrong claim. This arrangement satisfies the conditions of the puzzle. Unjustified conclusion
- Now, let's consider the other possible pair (9, 5). The product of 95 and 37 is 3515, ~~but subtracting 7 would~~ Lack of commonsense: 7 is no longer the central number, but 1 give us 3508, which does not have repetitions of a single figure.

Therefore, ~~the only valid arrangement is (5, 9, 3, 1, 7),~~ Unjustified claim where the number formed by the first pair (59) multiplied by the ~~number formed by the last pair (37),~~ Inconsistency: the number is 17 with the central number (3) subtracted, ~~gives us the number 2176.~~

Inconsistency: GPT has already stated that the result should be repetitions of the same figure Hence, ~~the solution to the puzzle is to arrange the cards as follows: 59 - 3 - 37.~~ Inconsistency: figure 1 is missing Wrong conclusion

Spoiler alert: Assessing performance

Answer	BARD	ChatGPT
Correct	5	7
Correct (but wrong justification)	13	2
Correct (with inconsistencies)	1	1
Correct (but unjustified)	-	6
Partial correct	-	1
Wrong	69	72
Wrong (lack of task understanding)	1	1
Wrong (claiming no solution)	11	2
Wrong (claiming not enough information)	-	3
Wrong and fuzzy	-	1
Wrong (no justification)	-	3
No solution only valid but inefficient algorithm	-	1
Admitting failure	-	1

Example (Solving strategies)

- Analysing all possibilities
- Backward reasoning
- Principle of inclusion-exclusion
- Trial and error
- Recursive approach
- Step by step
- Backtracking
- Euclidian algorithm
- Heron formula
- Assumption based

Quantifying logical faults

puzzle 1	Wrong	Wrong	20.36%
puzzle 2	Wrong	Correct	0.00%
puzzle 3	Wrong (said there is no solution)	Wrong	35.04%
puzzle 4	Wrong	Wrong	25.99%
puzzle 5	Wrong	Wrong	25.02%
puzzle 6	Wrong	Wrong	46.17%
puzzle 7	Wrong (did not understand task)	Correct	0.00%
puzzle 8	Wrong	Wrong	34.31%
puzzle 9	Wrong	Correct	0.00%
puzzle 10	Wrong (said there is no solution)	Wrong	26.16%
puzzle 11	Wrong (said there is no solution)	Wrong	26.41%
puzzle 12	Wrong	Wrong	9.61%
puzzle 13	Wrong (said there is no solution)	Wrong	37.57%
puzzle 14	Wrong	Wrong	47.03%
puzzle 15	Correct	Wrong	25.62%

$a \rightarrow b$

$b \rightarrow c$

$c \rightarrow d$

$d \rightarrow e$

$e \rightarrow \text{conclusion}$

conclusion

How much hallucination?

- on average, 26.03% from the generated text is a logical fault
- the quantity of false text is larger
- 698 logical faults (average 7 fallacies/puzzle)

1	inconsistency	22.35%	156	35	correct but unsupported conclusion	0.14%	1
2	implication does not hold	16.76%	117	36	false contradiction	0.14%	1
3	wrong conclusion	11.46%	80	37	fuzzy answer	0.14%	1
4	unsupported claim	9.17%	64	38	hallucination	0.14%	1
5	lack of commonsense	8.05%	56	39	identifying inconsistency and wrongly solving it	0.14%	1
6	bad arithmetic	5.16%	36	40	incomplete assignment	0.14%	1
7	wrong assumption	3.72%	26	41	incomplete claim	0.14%	1
8	unjustified claim	2.29%	16	42	incomplete conclusion	0.14%	1
9	too strong assumption	2.01%	14	43	incomplete exploration of the search space	0.14%	1
10	unjustified contradiction	1.43%	10	44	incomplete inference	0.14%	1
11	wrong justification	1.43%	10	45	incomplete justification	0.14%	1
12	unsupported conclusion	1.29%	9	46	incomplete reasoning	0.14%	1
13	lack of task understanding	1.00%	7	47	incomplete search space	0.14%	1
14	wrong claim	1.00%	7	48	incomplete search space exploration	0.14%	1
15	false contradiction	0.86%	6	49	incomplete solution by stoping before reaching	0.14%	1
16	lack of understanding	0.86%	6	50	inconsistency in the same sentence	0.14%	1
17	unjustified conclusion	0.86%	6	51	irrelevant justification	0.14%	1
18	unrelated justification	0.72%	5	52	loop	0.14%	1
19	logical fault	0.57%	4	53	partial solution	0.14%	1
20	bad justification	0.43%	3	54	the worlds collide	0.14%	1
21	incomplete analyse	0.43%	3	55	too restrictive domain	0.14%	1
22	incomplete solution	0.43%	3	56	too weak assumption	0.14%	1
23	missing task specification	0.43%	3	57	unrelated claim	0.14%	1
24	wrong solution	0.43%	3	58	unsupported premise	0.14%	1
25	fuzzy conclusion	0.29%	2	59	using info from previous puzzles	0.14%	1
26	lack of domain knowledge	0.29%	2	60	wrong claim based on previous errors	0.14%	1
27	language pattern	0.29%	2	61	wrong computation for favorable models	0.14%	1
28	too large domain	0.29%	2	62	wrong conclusion puzzle cannot be solved	0.14%	1
29	grammar patttern	0.29%	2	63	wrong modality	0.14%	1
30	bad interpretation in the last step	0.14%	1	64	wrong premise	0.14%	1
31	bad strategy	0.14%	1	65	wrong search space evaluation	0.14%	1
32	bad topology	0.14%	1	66	wrong strategy	0.14%	1
33	circular reasoning	0.14%	1	67	wrong undeterminacy	0.14%	1
34	correct but unjustified conclusion	0.14%	1				

Sample of correct puzzles

Puzzle 2. Logic equation

In this 4×4 logic equation you have to find unique integer values for the variables A , B , C , D - ranging from 1 to 4 - to make all the following statements true: $A + D = B + 4$, $B + D = A + 2$ (©Brainzilla - www.brainzilla.com)

Puzzle 7. Dividing the legacy

A man left 100 dollars to be divided between his two sons Alfred and Benjamin. If one-third of Alfred's legacy was taken from one-fourth of Benjamin's, the remainder would be 11 dollars. What was the amount of each legacy? (puzzle 15 from (Dudeney, 2016))

Puzzle 87. Latin square

Albert is a scientist that wants to test four different drugs (called A, B, C, and D) on four volunteers. He decides that every volunteer has to be tested with a different drug each week, but no two volunteers are allowed the same drug at the same time.

Inconsistencies and lack of commonsense

Inconsistencies: (156, 22.35%)

Lack of commonsense (56, 8.05%)

Example (Letter to parents: SEND + MOST = MONEY)

So, we have found the best case scenario: $SEND = 1529$, $MOST = 6735$, $MONEY = 8264$.
Implication does not hold Inconsistency: GPT already learned that MONEY has five digits

In the worst case scenario, you will receive \$89,675, and in the best case scenario, you will receive \$8,264.
Lack of commonsense

Example (Asking for ages of family members)

Therefore, the age of each person is: Father (F): 18 years Mother (M): 54 years John (J): 3 years Ben (B): 6 years Mary (Y): 9 years.
Lack of commonsense: a father of 18 cannot have a child of 9.
Lack of commonsense: a father of 18 and mother of 54 is rather exceptional.
Inconsistency: difference between father and mother (36) is not equal with the difference between john and mary (6) or ben and mary (3)

No justification

Puzzle 100. Self-counting sentence

You have to insert numbers in the blanks to make the following sentence true:
In this sentence, the number of occurrences of 0 is __, of 1 is __, of 2 is __, of 3 is __, of 4 is __, of 5 is __, of 6 is __, of 7 is __, of 8 is __, and of 9 is __.
There are exactly two solutions. (puzzle created by Douglas Hofstadter [Hofstadter \(1985\)](#))

Solution by ChatGPT

Wrong (no justification)

The sentence can be made true in the following two ways:
Solution 1: "In this sentence, the number of occurrences of 0 is 2, of 1 is 2, of 2 is 1, of 3 is 1, of 4 is 1, of 5 is 1, of 6 is 1, of 7 is 1, of 8 is 1, and of 9 is 1." Unjustified conclusion Wrong conclusion
Solution 2: "In this sentence, the number of occurrences of 0 is 4, of 1 is 1, of 2 is 1, of 3 is 1, of 4 is 2, of 5 is 1, of 6 is 1, of 7 is 1, of 8 is 0, and of 9 is 1." Unjustified conclusion Wrong conclusion

Lack of task understanding

Puzzle 89. Fancy queens

I have placed a queen in one of the white squares of the 8x8 chessboard. The queen is placed on the position A2. Place 7 more queens in white squares so that no 2 of the 8 queens are in line horizontally, vertically, or diagonally (adapted from puzzle 113 from [Kordemsky \(1992\)](#)).

Solution by ChatGPT

Wrong

To solve this puzzle, you can place the ~~remaining~~ queens Lack of task understanding: there is no queen on A2 on the chessboard in such a way that no two queens threaten each other horizontally, vertically, or diagonally. ~~Here's a configuration that satisfies these conditions:~~ Wrong conclusion

```

Q . . . . .
. . Q . . .
. . . Q . .
. Q . . . .
. . . . Q .
. . . Q . .
. . . . . Q
. . . . . Q

```

In this configuration, each row, column, and diagonal contains only one queen, and no two queens threaten each other.

**Solution by ChatGPT****Correct, but unsupported**

Let's analyze the possibilities: If the sign on the first room is true, then a tiger is in that room. This means the sign on the second room must be false Implication does not hold, and the sign on the third room must be true Unsupported claim. However, this creates a contradiction because if the sign on the third room is true, then a tiger is in room 2, which means the sign on the second room should also be true Implication does not hold (since it says a lady is in that room) Unrelated justification. So this possibility is not valid.

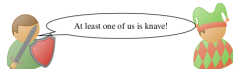
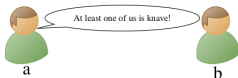
If the sign on the first room is false, then a lady is in that room. This means the sign on the second room must be true Implication does not hold, and the sign on the third room must be false. This is a valid possibility because it satisfies the condition that at most one of the three signs is true.

Based on the valid possibility Wrong justification: possibility does imply necessity, we can determine that the lady is in the first room Unsupported conclusion. Therefore, you should open the first door to find the lady.

Traces of the past

Puzzle 62. At least one of us is a knave

On the island of knights and knaves, knights always tell the truth, while knaves always lie. You are approached by two people. The first one says: “At least one of us is a knave”. What are they actually? (puzzle 28 from [Smullyan \(2011\)](#))



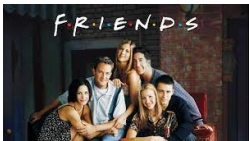
Solution by ChatGPT

Wrong

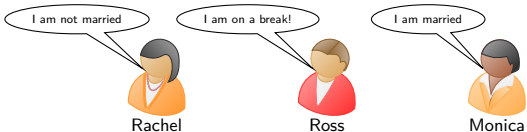
If we assume that the first person is a knight, then what they say must be true. However, this leads to a contradiction. If the first person is a knight and they are telling the truth, then both of them can't be knaves **as stated**.
Language pattern: maybe GPT uses it to avoid gender bias
Using info from previous puzzles: this contradicts their statement "At least one of us is a knave."
False contradiction: the statement "At least one of us is a knave" does not contradict the statement "Both of us are knaves"

If we assume that the first person is a knave, then what they say must be false. This is consistent because a knave lying is expected behavior. If the first person is a knave and they are lying, then it's possible that neither of them is a knave.
Wrong modality: It's should
~~Therefore, the only consistent interpretation is that the first person is a knave and the second person is a knight.~~
Wrong conclusion This way, the statement "At least one of us is a knave" made by the knave is false,
Inconsistency: the statement is true and the second person (a knight) tells the truth.
Lack of understanding: the second person says nothing

The worlds collide



Example
*There are three friends staying on the couch in Central Perk: Rachel, Ross, and Monica. Monica is looking at Ross. Ross is looking at Rachel. Monica is married; Rachel is not. **Is a married person looking at an unmarried person?***



Solution by ChatGPT

Correct (but wrong justification)

Yes, a married person is looking at an unmarried person. ~~Ross is married to Rachel~~ The worlds collide: GPT uses information from Friends TV series and he is looking at her. ~~Since Rachel is unmarried,~~ Inconsistency: GPT has just stated that Ross is married to Rachel this scenario fits the given information.

Admitting failure - the most interesting answer

Puzzle 24. An unusual number

Find a six-digit number which, when multiplied by an integer between 2 and 9 inclusive, gives the original six-digit number with its digits reversed. Thus, if the original number was 123,456, and the chosen integer is 8, then $123,456 \times 8$ should equal 654,321, which, of course, it doesn't. However, it is possible to find more than one solution to this problem, but I'll accept anyone that meets the required condition (puzzle 34 from [Clessa 1996](#))

Solution by ChatGPT

To find a six-digit number that, when multiplied by an integer between 2 and 9 inclusive, gives the original number with its digits reversed, we can use trial and error.

Solving strategy: Trial and error: for GPT this is always incomplete.

Solving strategy: Wrong since it has to try all numbers of 6 digits.

to check different numbers. We need to find a number that, when multiplied by any integer between 2 and 9, gives a result that has the same digits as the original number but in reverse order.

Task understanding: It seems so

Let's start with a systematic approach: Start with a six-digit number in the form of ABCDEF. Check all possible multipliers from 2 to 9 (inclusive) and calculate the product. Check if the product has the same digits as the original number ABCDEF but in reverse order.

Solving strategy: Brute force: it seems like an algorithm

Let's illustrate this process with an example: Original number: ABCDEF Let's assume $ABCDEF = 123456$ as an example. Now, we'll check all possible multipliers from 2 to 9.

When multiplied by 2: $123456 * 2 = 246912$ (not a solution).

When multiplied by 3: $123456 * 3 = 370368$ (not a solution).

When multiplied by 4: $123456 * 4 = 493824$ (not a solution).

When multiplied by 5: $123456 * 5 = 617280$ (not a solution).

When multiplied by 6: $123456 * 6 = 740736$ (not a solution).

When multiplied by 7: $123456 * 7 = 864192$ (not a solution).

When multiplied by 8: $123456 * 8 = 987648$ (not a solution).

When multiplied by 9: $123456 * 9 = 1111104$ (not a solution).

Admitting failure

Admitting failure - the most interesting answer

Let's try another number: Original number: ABCDEF Let's assume ABCDEF = 142857 as another example. Now, we'll check all possible multipliers from 2 to 9.

When multiplied by 2: $142857 * 2 = 285714$ (not a solution).

When multiplied by 3: $142857 * 3 = 428571$ (not a solution).

When multiplied by 4: $142857 * 4 = 571428$ (not a solution).

When multiplied by 5: $142857 * 5 = 714285$ (not a solution).

When multiplied by 6: $142857 * 6 = 857142$ (not a solution).

When multiplied by 7: $142857 * 7 = 999999$ (not a solution).

When multiplied by 8: $142857 * 8 = 1142856$ (not a solution).

When multiplied by 9: $142857 * 9 = 1285713$ (not a solution).

None of the products have the same digits as the original number in reverse order for the example ABCDEF = 142857.

It is important to note that there are six-digit numbers that satisfy this condition, Solving strategy: Admitting failure may be trigger by the input sentence: It is possible to find more than one solution"

but finding them requires a more systematic approach or potentially using a computer program to check all possible combinations. In this particular example, we didn't find a solution for ABCDEF = 123456 or ABCDEF = 142857. Solving strategy: Interesting formulation: we didn't find a solution.

Abusing epistemic notation: $K_{ChatGPT}(\neg K_{ChatGPT} answer) \rightarrow K_{ChatGPT} answer$

Abusing temporal logic: $AG(K_{ChatGPT}(\neg K_{ChatGPT} answer)) \rightarrow EF(K_{ChatGPT} answer)$

Lost in translation

Solving logical puzzles with LLMs is not a good use case.

Example (But, LLMs master language translations)

- 1 NL to First Order Logic - for **human robot interaction** (kitchen robot)
- 2 NL to SUO-KIF - for **detecting misinformation** (diet domain)
- 3 NL to OWL - for supporting **ontology engineering**
- 4 NL to First Order Logic - for **reasoning** tasks (puzzle domain)



Interpretation models

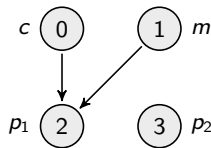
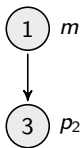
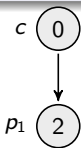
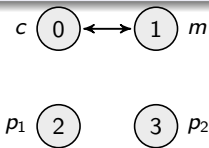


Monica is in love and Chandler is in love.

$\exists x, \text{love}(\text{monica}, x) \wedge \exists x, \text{love}(\text{chandler}, x)$

How many models are (e.g. MACE4)?

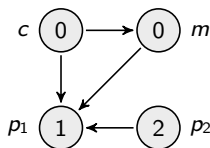
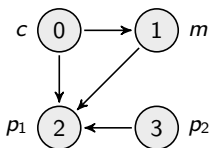
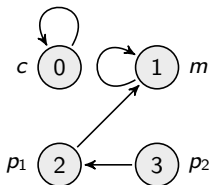
```
assign(max_models, -1).
assign(domain_size, 4).
formulas(assumptions).
    exists x love(chandler,x).
    exists x love(monica,x).
end_of_list.
```



```
===== STATISTICS =====
For domain size 4.
Current CPU time: 0.00 seconds (total CPU time: 5.66 seconds)
Ground clauses: seen=2, kept=2.
Selections=278522, assignments=557049, propagations=18, current
Rewrite_terms=23, rewrite_bools=20, indexes=18.
Rules_from_neg_clauses=0, cross_offs=0.

===== end of statistics =====
User_CPU=5.66, System_CPU=10.18, Wall_clock=25.
Exiting with 278528 models.
----- process 4061 exit (all_models) -----
Process 4061 exit (all_models) Sun Jul 28 11:51:49 2019
```


Reducing 278,528 models

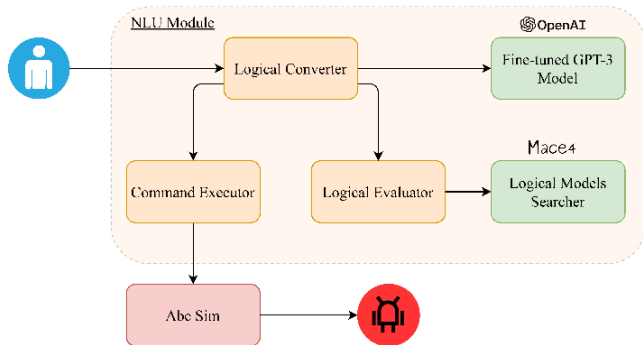


- ① UNA: *chandler* \neq *monica* **163,840** models
- ② Assume love is not narcissistic: $\forall x, \neg \text{love}(x, x)$. (**5,120**)
- ③ Assume someone can love only one person at a time:
 $\text{love}(x, y) \wedge \text{love}(x, z) \rightarrow y = z$. (**80**)
- ④ Remove isomorphic interpretations (**74**)
- ⑤ 2 Skolem consts - assume no interest in love relations between them (**17**)

Remarks

- ① order of reductions is computationally relevant (186,976 models, 2h))
- ② which domain knowledge to add is subject to interpretation

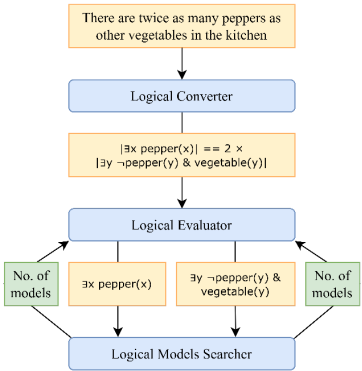
Natural language quantifiers for human-robot interaction ¹



Command type	Example
Quantifiers	“Fetch <i>all</i> green peppers” “Cut <i>several</i> bananas with a knife” “Cover <i>3</i> trays with paper”
Referred objects	“Move contents of <i>MediumBowl1</i> to <i>MediumBowl2</i> ”
Quantifiers and referred objects	“Next cut <i>1</i> mango using cooking knife <i>Knife1</i> ”

¹S. Morar, A. Groza, M. Pomarlan, Natural language quantifiers for human-robot interaction, AIC, Bremen, Germany 14-15 September 2023

Query	FOL with cardinality
All objects are boxes	$\forall x \text{ object}(x) \rightarrow \text{box}(x)$
No object is a box	$\neg \exists x \text{ object}(x) \wedge \text{box}(x)$
There is a box	$\exists x \text{ object}(x) \wedge \text{box}(x)$
There are at least two boxes	$ \exists x \text{ box}(x) \geq 2$
There are exactly two boxes	$ \exists x \text{ box}(x) == 2$
There are more boxes than tools	$ \exists x \text{ box}(x) > \exists y \text{ tool}(y) $
Most objects are boxes	$ \exists x \text{ box}(x) \ \& \ \text{object}(x) > \exists y \neg \text{box}(y) \ \& \ \text{object}(y) $
There are twice as many boxes as other objects	$ \exists x \text{ box}(x) == 2 \times \exists y \neg \text{box}(y) \ \& \ \text{object}(y) $
There are many boxes	$ \exists x \text{ box}(x) \geq \textit{threshold}$
How many boxes are there?	$ \exists x \text{ box}(x) $



```

formulas (expressions) .
  exists x (ingredient(x)) .
end_of_list .

```

Listing 7: Sample content for file "expression.in"

Generating interpretation models

```
assign(domain_size, 5).

list(distinct).
  [Robot1, Tomato1, Tomato2, Whisk1, CookingKnife1].
end_of_list.

formulas(sensors).
  robot(Robot1). tomato(Tomato1). tomato(Tomato2).
  whisk(Whisk1). cookingKnife(CookingKnife1).
end_of_list.
```

Listing 5: Sample content for file “sensors.in”

```
formulas(background_knowledge_classification).
  tomato(x) -> ingredient(x).
  cookingKnife(x) -> kitchenTool(x). whisk(x) -> kitchenTool(x).
end_of_list.

formulas(background_knowledge_distinction).
  ingredient(x) | kitchenTool(x) -> -robot(x).
  robot(x) | kitchenTool(x) -> -ingredient(x).
  robot(x) | ingredient(x) -> -kitchenTool(x).
  cookingKnife(x) -> -whisk(x).
end_of_list.

formulas(background_knowledge_commands).
  robot(x) & (ingredient(y) | kitchenTool(y)) -> fetch(x, y).
  -robot(x) -> -fetch(x, y).
  -ingredient(y) & -kitchenTool(y) -> -fetch(x, y).
end_of_list.
```

Dataset for experiments

Table 4

Dataset distribution

Pair type	Train	Dev	Test
Command	240	80	80
Query	96	32	32
Invalid	60	20	20
Total	396 (60%)	132 (20%)	132 (20%)

Table 5

Command distribution in the dataset

Command	Train	Dev	Test
fetch	59	19	19
cut	38	12	12
bake	36	12	12
line	36	12	12
mix	26	9	9
transfer	16	6	6
sprinkle	16	6	6
shape	13	4	4
Total	240	80	80

Table 6

Query distribution by quantifier

Query	Train	Dev	Test
most/majority of	6	3	3
more than	6	2	3
less than	5	1	2
at most	4	1	1
at least	3	1	1
exactly/only	7	2	1
n	3	1	2
n times more	4	2	1
between k_1 and k_2	4	1	1
many/a lot	4	2	1
several	3	1	1
a few/few	5	2	2
a couple	3	1	1
some	4	1	1
how many/count	7	2	2
half	3	1	2
no/none	4	2	1
all/every	8	3	3
dozen/half a dozen	4	1	1
<i>combinations</i>	9	2	2
Total	96	32	32

Results - 132 testing examples

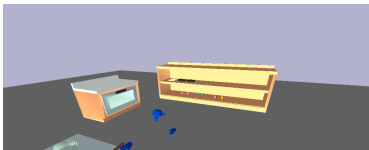
- 93 translations (70,45%) identical to the expected ones;
- 7 translations (5,30%) small variations, but logically equivalent
- 32 translations (24,24%) wrong, leading to a different interpretation

```
# Expected
{'type': 'command', 'expressions': [['| exists x2 (whisk(x2)) .| >= 1']],
 'commands': ['robot(x0) & bowl(Bowl1) & whisk(x2) -> mix(x0, Bowl1,
x2) . ']}
# Generated
{'type': 'command', 'expressions': [['| exists x1 (whisk(x1)) .| >= 1']],
 'commands': ['robot(x0) & bowl(Bowl1) & whisk(x1) -> mix(x0, Bowl1,
x1) . ']}
```

Listing 8: A different but correct translation for the prompt: “Blend the contents of the bowl Bowl1 using a whisk”

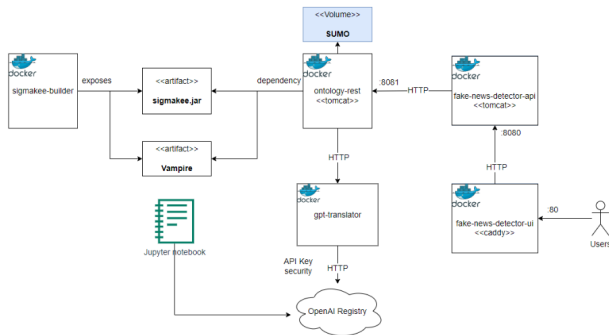
```
# Expected
{'type': 'query', 'expressions': ['all x0 (pepper(x0) -> -redPepper(x0)) . ']}
# Generated
{'type': 'query', 'expressions': ['all x0 (pepper(x0) -> redPepper(x0)) . ']}
```

Listing 9: Incorrect translation for the prompt “All peppers are not red chili peppers”



Cross-validation of Answers with SUMO and GPT²

Sentence	Fact check
Some sugar causes obesity	True
Calcium strengthens bones	True
Protein builds muscles	True
Broccoli contains vitamins	True
Fruits and vegetables are healthy	True
Salmon contains Omega3	True
Detox diets cleanse the blood	False
Some lipids are unhealthy	False
Eggs raises cholesterol	False
All salts are unhealthy	False
Some lipids are healthy	True
Coffee dehydrates you	Unknown
Athletes consume more protein	Unknown
Some detox diets cleanse the body	True



```
{text : "Deserts are dry"  
formal: "If ?X is an instance of desert , then ?X has the attribute dry"  
kif : "(=> (instance ?X Desert) (attribute ?X Dry))"}
```

Listing 1: Training example for attribute

```
{text : "Apples and bananas are fruits"  
formal: "Apple and banana are subclasses of fruit",  
kif : "(and (subclass Apple Fruit) (subclass Banana Fruit))"}
```

Listing 2: Training example for subclass relation

```
{text : "Broccoli contains vitamins."
 formal: "If ?B is an instance of broccoli , then there exists ?V such
         that ?V is an instance of vitamin and ?V is part of ?B."
 kif    : "(=> (instance ?B Broccoli) (exists (?V)
              (and (instance ?V Vitamin) (part ?V ?B))))" }
```

Listing 3: Training example for contains/part relation

```
{text : "Parasites damage cells."
 formal: "If ?P is an instance of parasite , then there exists ?D such
         that ?D is an instance of damaging , and ?P is the agent of
         ?D and cell is the patient of ?D."
 kif    : "(=> (instance ?P Parasite) (exists (?D)
              (and
                (instance ?D Destroying)
                (agent ?D ?P)
                (patient ?D Nutrient))))" }
```

Listing 4: Training example for agent-patient relation

Type	Training	Testing
Agent-Patient	106	21
Attribute	150	30
Subclass	77	15
Contains-part	85	17
Total	418	83

NL to KIFNL 96%
 KIFNL to KIF 82%
 76/83 correct translations (92%)

Natural Language to OWL³

The screenshot displays two instances of the GPT Ontology Augmenter interface. Each instance consists of a left sidebar with class and individual hierarchies, a central area for property assertions, and a right panel for text input and resulting OWL code.

Top Instance:

- Class hierarchy:** owl:Thing (parent), girl (child).
- Property assertions:** Object property assertions (empty), Data property assertions (empty).
- Individuals:** girl (2) (parent), Lana (child), Anna (child).
- Text input:** "Anna and Lana are girls."
- Generated OWL:**

```
Declaration(Class(:girl))
Declaration(NamedIndividual(:Anna))
Declaration(NamedIndividual(:Lana))
ClassAssertion(:girl :Anna)
ClassAssertion(:girl :Lana)
```

Bottom Instance:

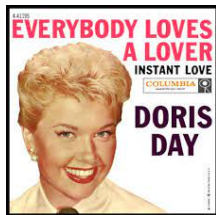
- Class hierarchy:** owl:topObjectProperty (parent), has_cousin (child), has_sister (child).
- Property assertions:** Object property assertions: has_cousin (Nola), has_sister (Lana).
- Individuals:** Nola (parent), girl (2) (parent), Anna (child), Lana (child).
- Text input:** "Nola and Anna are each other's cousins."
- Generated OWL:**

```
Declaration(ObjectProperty(:has_cousin))
Declaration(NamedIndividual(:Anna))
Declaration(NamedIndividual(:Nola))
ObjectPropertyAssertion(:has_cousin :Anna :Nola)
ObjectPropertyAssertion(:has_cousin :Nola :Anna)
```

Assume Phoebe sings one of the Doris Day songs:



Everybody loves a lover
I'm a lover, everybody loves me
... And I love everybody,
since I fell in love with you



Prompt: "Translate into First Order Logic"

p_1	"Everybody loves a lover"	$\forall x (lover(x) \rightarrow \forall y loves(y, x))$
p_2	"I feel in love with you":	$loves(I, You)$
p_3	"I am a lover"	$lover(I)$
c_2	"Everybody loves me"	$\forall x loves(x, I)$
c_3	"I love everybody"	$\forall x loves(I, x)$

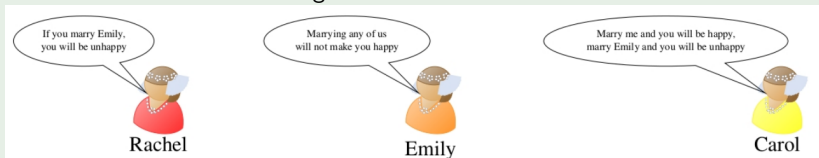
ChatGPT "proves" c_1 but not c_2 .

Based on three premises p_1 ="Everybody loves a lover", p_2 ="I am a lover" and p_3 ="I feel in love with you" can you prove c_2 ="I love everybody"?

ChatGPT + Prover9 prove both c_1 and c_2 .

Example (Ross trilemma)

Ross is in situation to choose marrying one of Rachel, Emily or Carol. Each woman sent Ross some messages.



Given that Ross knows that all the women lie, will Ross be marry or not. Will he be happy or not? Will he marry Rachel or not?

Close the world to four individuals - $domainsize = 4$

Avoid isomorphic models: $Ross = 0$, $Rachel = 1$, $Emily = 2$, $Carol = 3$

$$m(Rachel) \leftrightarrow (marry(Emily) \rightarrow \neg happyRoss)$$

$$m(Emily) \leftrightarrow (\forall x \text{ marry}(x) \rightarrow \neg happyRoss)$$

$$m(Carol) \leftrightarrow ((marry(Carol) \rightarrow \neg happyRoss) \wedge (marry(Emily) \rightarrow happyRoss))$$

All messages are false: $\neg m(Rachel) \wedge \neg m(Emily) \wedge \neg m(Carol)$

Helping the solver vs. helping ChatGPT

- Ross should be married: $\exists x \text{ marry}(x)$
- Marry only one: $\forall x \forall y (\text{marry}(x) \wedge \text{marry}(y) \rightarrow x = y)$
- Ross cannot married to himself: $\exists x x \neq Ross \wedge \text{marry}(x)$



Students find the single model. In this model Ross is happy (by marrying Emily).

Prompt: Solve the following puzzle

Since we know that all the women lie, none of their statements can be true.[...] If Ross marries Rachel, [...] If Ross marries Emily, [...] If Ross marries Carol, [...] Since there is no consistent information among the statements, Ross's decision cannot be based on these statements alone. Therefore, Ross cannot make a logical decision based on the information provided. Whether Ross marries or not and whether he is happy or not will be uncertain in this scenario.

Consider that Ross should marry at least one woman.

...Since Ross must marry at least one woman, he has two choices: Rachel and Carol. .. Ross should choose based on his own feelings and preferences. Ross's happiness in marriage will depend on his compatibility with the chosen woman, which cannot be determined from the information provided.

Does Ross can marry more than one woman?

[...] based on the information provided and standard assumptions about marriage, Ross can only marry one woman.

Does Ross can marry to himself?

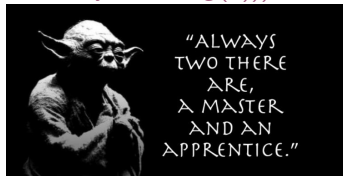
While self-love and self-acceptance are important, they are typically not expressed through the institution of marriage to oneself. Instead, individuals can practice self-care, self-acceptance, and self-improvement to promote their

$\neg \exists x \text{ Unreal}(x) \wedge \text{Exists}(x)$

$\forall x \exists y (\text{Master}(x) \wedge \text{Apprentice}(y) \wedge x \neq y) \wedge \forall x \forall z (\text{Master}(x) \wedge \text{Apprentice}(z) \rightarrow x \neq z)$

$\forall x (\text{Wise}(x) \rightarrow (\text{Speaks}(x) \wedge H(x)))$ where $H(x) = \text{HasSomethingToSay}(x)$

$\forall x (\text{Fool}(x) \rightarrow (\text{Speaks}(x) \wedge H(x)))$ where $H(x) = \text{HasToSaySomething}(x)$



Wise men speak because they have something to say; Fools because they have to say something.
- Plato



There are three friends staying on the couch in Central Perk: Rachel, Ross, and Monica.

$\forall x (F(x) \wedge S(x)) \rightarrow (x = \text{Rachel} \vee x = \text{Ross} \vee x = \text{Monica})$

Monica is looking at Ross: $L(\text{Monica}, \text{Ross})$

Ross is looking at Rachel. $L(\text{Ross}, \text{Rachel})$

Monica is married. $M(\text{Monica})$

Rachel is not married. $\neg M(\text{Rachel})$

Theorem to prove: $\exists x \exists y (M(x) \wedge U(y) \wedge L(x, y))$

ChatGPT performance on 100 puzzles

- 7 correct puzzles
- on average, 26.03% from the generated text is a logical fault
- 698 logical faults (average 7 fallacies/puzzle)

Lost in translation with GPT models

- First Order Logic - for human robot interaction (kitchen robot)
- SUO-KIF - for detecting misinformation (diet domain)
- OWL - for supporting ontology engineering (family ontology)
- First Order Logic - for reasoning tasks (puzzle domain)

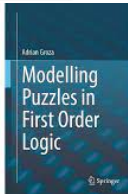
Groza, A.: *Modelling Puzzles in First Order Logic*. Springer (2021)

<https://users.utcluj.ro/~agroza/puzzles/maloga/codes.html>

https://users.utcluj.ro/~agroza/puzzles/maloga/chatGPT_puzzles.pdf

<https://users.utcluj.ro/~agroza/puzzles/maloga/100puzzles.txt>

Adrian.Groza@cs.utcluj.ro



$\forall x (\text{participant}(x, WG4 - WG5) \rightarrow \text{thank}(I, x))$