MIRAGE： Evaluating and Explaining Inductive Reasoning Process in Language Models

Jiachun Li[1,2], Pengfei Cao[1,2], Chenhao Wang[1,2], Zhuoran Jin[1,2], Yubo Chen[1,2] , Kang Liu[1,2], Jun Zhao[1,2]

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences
[2] National Laboratory of Pattern Recognition, Institute of Automation, CAS

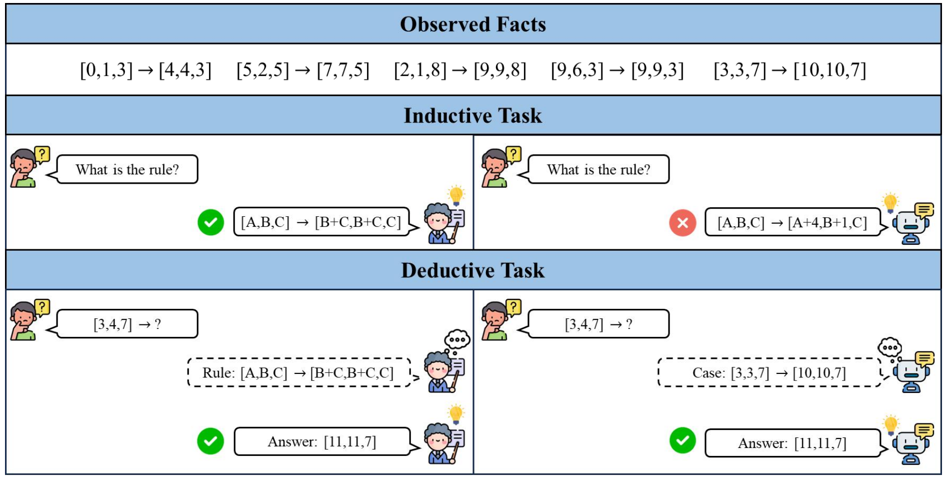
中国科学院
CHINESE ACADEMY OF SCIENCES

中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

Background

Inductive Reasoning

- Inductive reasoning refers to the ability of intelligent entities to **generalize rules** from limited observations and **apply them to unseen examples**.
  - Rule Induction (Inductive)
    - E.g. The sum of the interior angles of a triangle is 180° -> the sum of the interior angles of an n-gon is (n-2) * 180°
  - Example Inference (Deductive)
    - E.g. The sum of the interior angles of a triangle is 180° -> the sum of the interior angles of a pentagon is 540°



Rule-based Neighbor-based

Motivation

Limitations of Related Works

- Previous works lack comprehensive evaluation, most works have only one evaluation task:
  - Inductive task on collected rules
  - Deductive task on specific test samples

- Previous works lack flexible test data, hindering a deeper analysis of the model's reasoning mechanisms.
  - Fixed data distribution
  - Fixed test difficulty
  - Fixed task form

Dataset Construction

Dataset Construction Framework

Rule Generation

- Construct different abstract rules based on basic transformation functions

Fact Generation

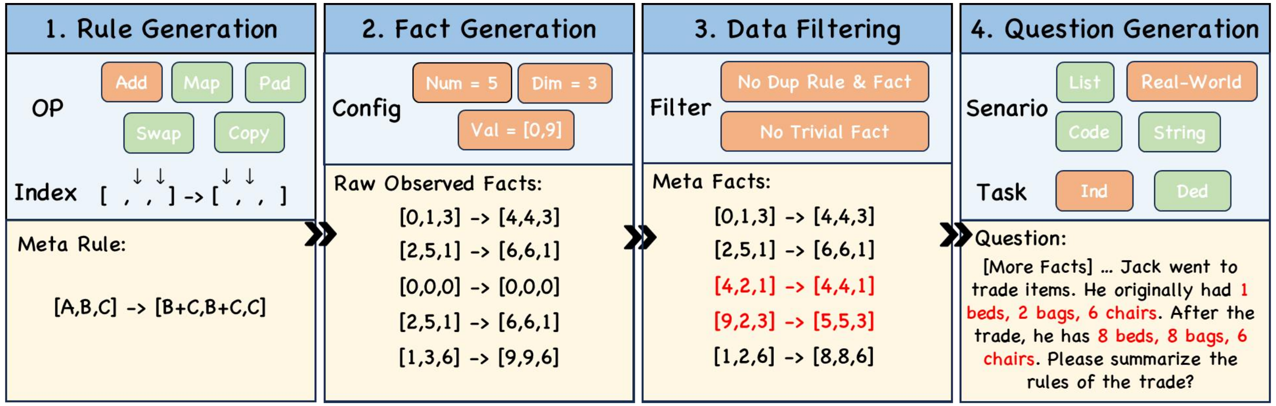
- Instantiate the abstract rules to generate any observations and test metadata

Data Filtering

- Filter low-quality and noisy data based on certain heuristic rules

Problem Generation

- Place the metadata into different reasoning scenarios to generate specific test questions in various forms

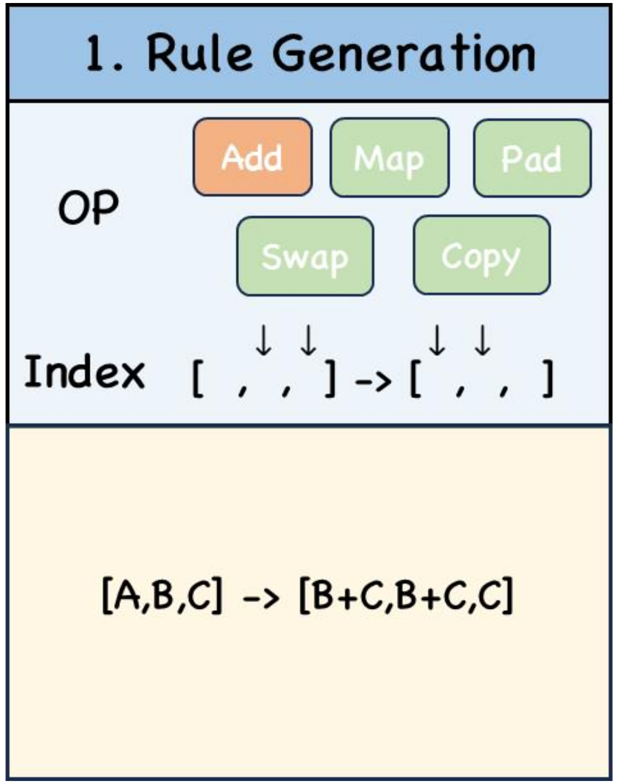


Dataset Construction

Rule Generation

- Formalize the rule as functions f in vector space, including five operations:
  - Add: Add certain elements together ([x,y,z] → [x,x+y,z])
  - Copy: Copy certain elements to other positions ([x,y,z] → [x,x+y,z])
  - Map: Apply linear transformations to certain elements ([x,y,z] → [x, ky+b, z])
  - Pad: Fill certain elements' positions with constants ([x,y,z] → [x, 1, z])
  - Swap: Swap the positions of elements ([x,y,z] → [y,x,z])

- By mixing the rules generated from all the above operations, we form the rule library for the dataset.



Dataset Construction

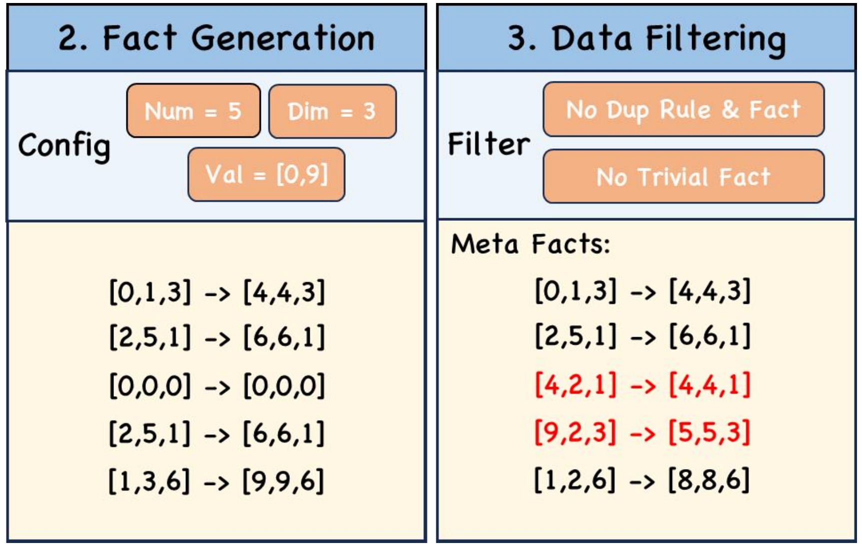
Fact Generation

- For each rule f , we randomly generate input vectors x and use f to generate the corresponding output y, hence constructing a collection of observation and test examples (x, y) as the fact library.

- We control the test difficulty by adjusting the count N of the observation examples and the dimensionality D of the vectors.

Data Filtering

- Use heuristic rules to filter the generated data
  - E.g. Remove duplicate rules and facts.

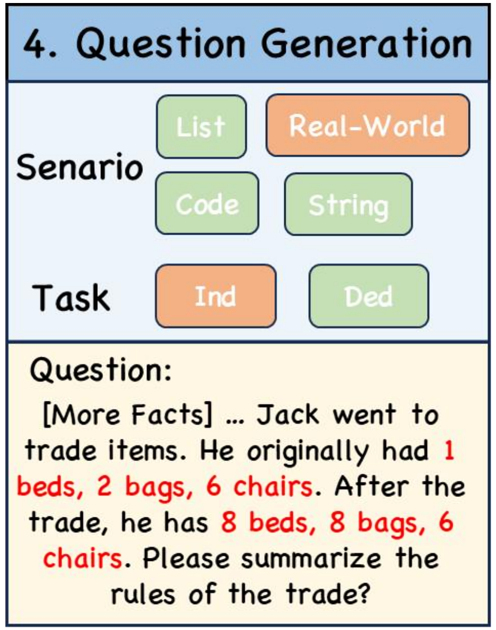


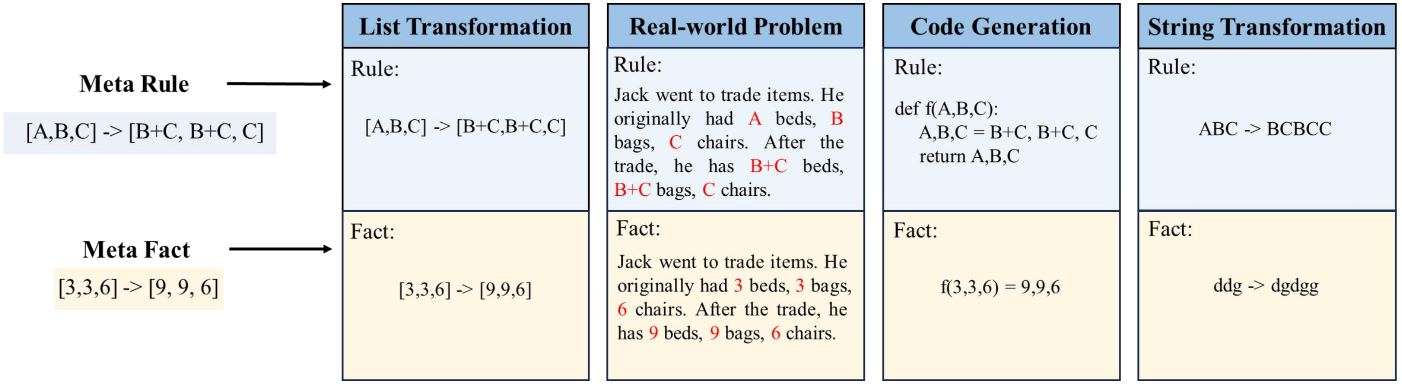
Dataset Construction

Question Generation

- Apply the generated rules and facts to different scenarios, including:
  - List Transformation (LT): Perform inductive operations on lists;
  - Real-World Problems (RP): Problems described in natural language;
  - Code Generation (CG): Generate corresponding Python functions based on input facts;
  - String Transformation (ST): Replace numeric operations with string operations.

- Design two tasks to evaluate the entire inductive reasoning process:
  - Rule Induction (RI): The model is required to correctly generate abstract rules;
  - Example Inference (EI): The model is required to correctly answer the test questions.





Model Evaluation

Overall Results

- The inductive reasoning ability of LLMs is relatively weak, especially in real-world scenarios (RP).
  - GPT-4o achieves only 0.16 and 0.17 accuracy on RP problems (D = 8).

- The deductive ability (EI) of LLMs is generally stronger than their inductive ability (RI).
  - Unlike humans, the inductive reasoning performance of LLMs does not ultimately depend on whether they can abstract the correct rules.

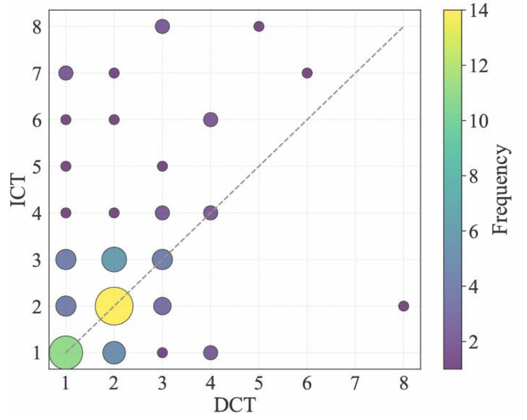
Model	Task	D=3				D=5				D=8			
		LT	RP	CG	ST	LT	RP	CG	ST	LT	RP	CG	ST
Llama2-13B	RI	0.01	0.00	0.00	0.03	0.01	0.01	0.00	0.21	0.00	0.01	0.00	0.10
	EI	0.26	0.11	0.25	0.22	0.13	0.03	0.14	0.25	0.06	0.01	0.06	0.19
Llama3-8B	RI	0.15	0.11	0.19	0.19	0.23	0.04	0.14	0.22	0.16	0.02	0.08	0.21
	EI	0.30	0.15	0.25	0.25	0.20	0.12	0.25	0.29	0.09	0.11	0.16	0.24
GPT-4o	RI	0.41	0.32	0.38	0.32	0.35	0.21	0.44	0.30	0.33	**0.16**	0.41	0.24
	EI	0.68	0.37	0.61	0.56	0.58	0.25	0.64	0.39	0.42	0.17	0.49	0.29
GPT-4	RI	**0.47**	0.29	**0.41**	0.28	**0.58**	**0.22**	**0.56**	0.27	**0.46**	0.15	**0.45**	0.23
	EI	0.68	0.37	0.61	0.57	0.63	0.29	0.71	0.44	0.42	0.21	0.64	0.30
Claude-3.5	RI	0.44	**0.35**	0.34	**0.46**	0.22	0.20	0.38	**0.33**	0.24	0.13	0.38	**0.26**
	EI	0.79	0.45	0.62	0.58	0.65	0.33	0.76	0.45	0.46	0.24	0.59	0.30

Reasoning Mechanism Analysis

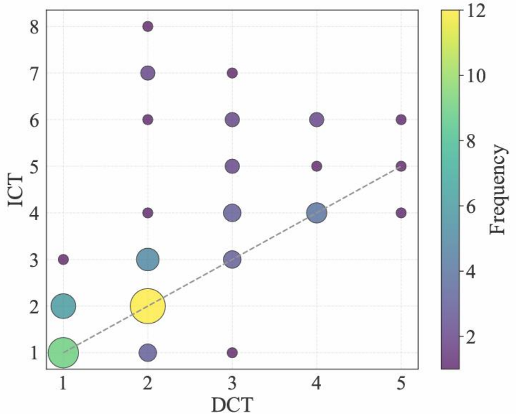
Langauge Models are Poor Rule-based Reasoners

- Compute the minimum number of observed examples required to correctly induce the rules (ICT) and answer the test questions (DCT).
  - If reasoning is based on rules, ICT should be less than DCT for most examples

- The inductive reasoning of LLMs on new examples **does not rely on successful rule induction**.



(a) GPT-4o



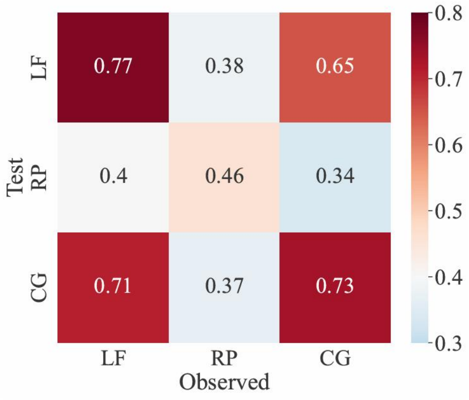
(b) Claude-3.5

Reasoning Mechanism Analysis

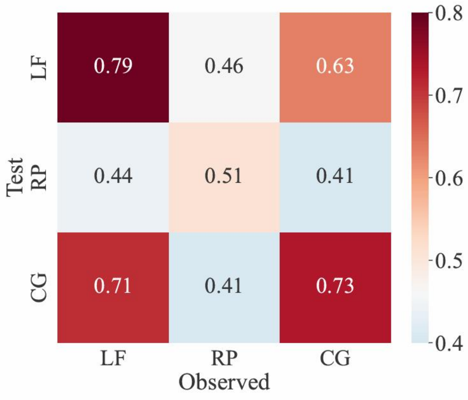
Inductive Reasoning across Different Task Format

- For a fixed observed facts, test the model's performance under other task formats.

- LLM's inductive reasoning rely on the structural similarity between observed and test examples



(a) GPT-4o



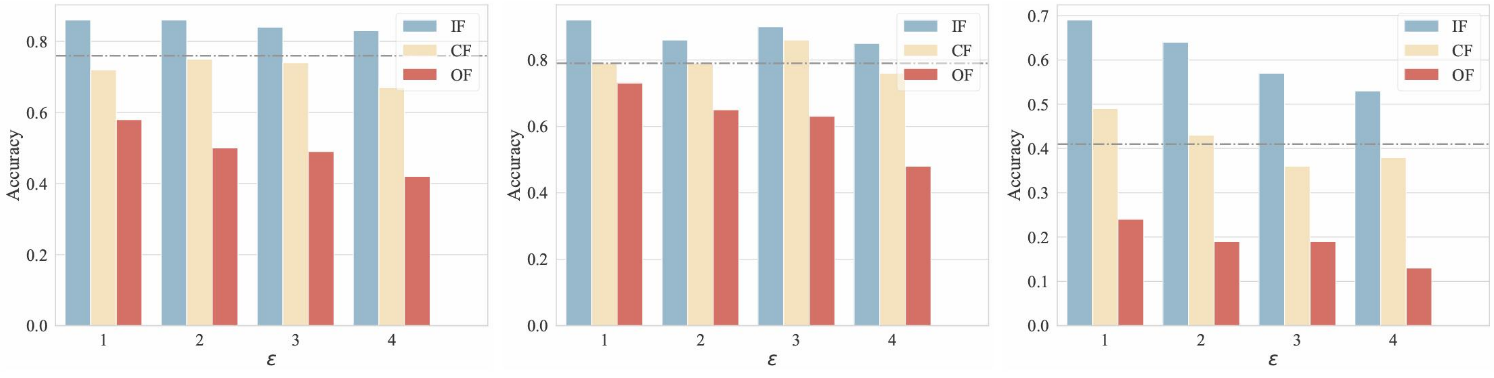
(b) Claude-3.5

Reasoning Mechanism Analysis

Langauge Models are Good Neighbor-based Reasoners

- We categorize observed facts into three types based on their distance to the test input:
  - **In-neighborhood Fact (IF):** All corresponding components are in the neighborhood
  - **Cross-neighborhood Fact (CF):** Some of the components are in the neighborhood
  - **Out-neighborhood Fact (OF):** None of the components are in the neighborhood

- The closer the observed examples are to the test examples, the better the reasoning performance

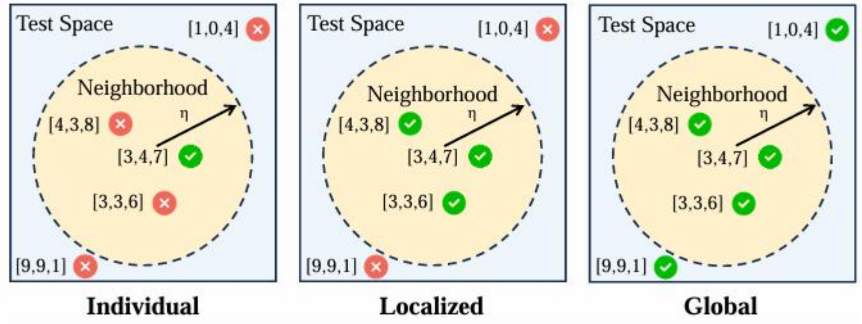


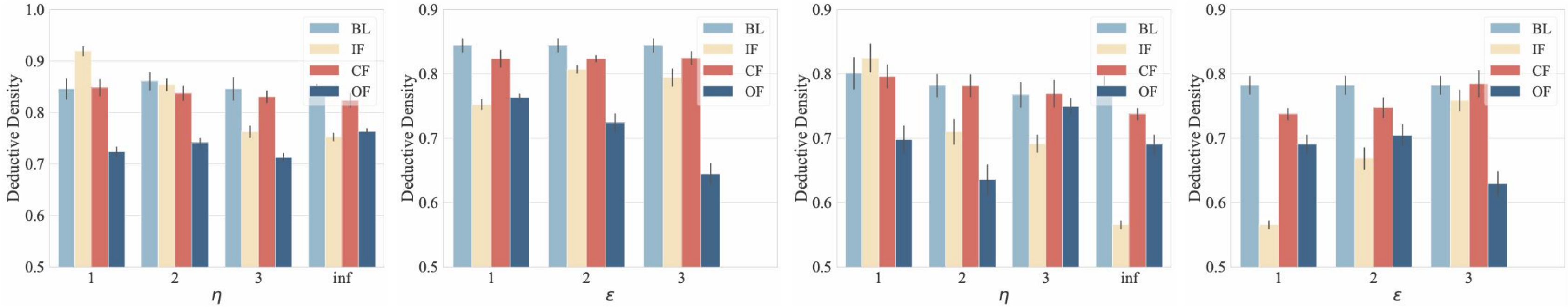
(a) GPT-4o (b) Claude-3.5 (c) Llama3-8B

Reasoning Mechanism Analysis

Effective Scope of Neighbor-based Reasoning

- Sample multiple additional test examples within a certain radius, calculate the average accuracy to evaluate the effective scope;
- The scope of neighbor-based reasoning is local;
- The more dispersed of the neighbor facts distribution, the larger the effective scope.





(a) $D = 3, \epsilon = 1$ (b) $D = 3, \eta = \text{inf}$ (c) $D = 5, \epsilon = 1$ (d) $D = 5, \eta = \text{inf}$

Thanks