Language Models as Inductive Reasoners

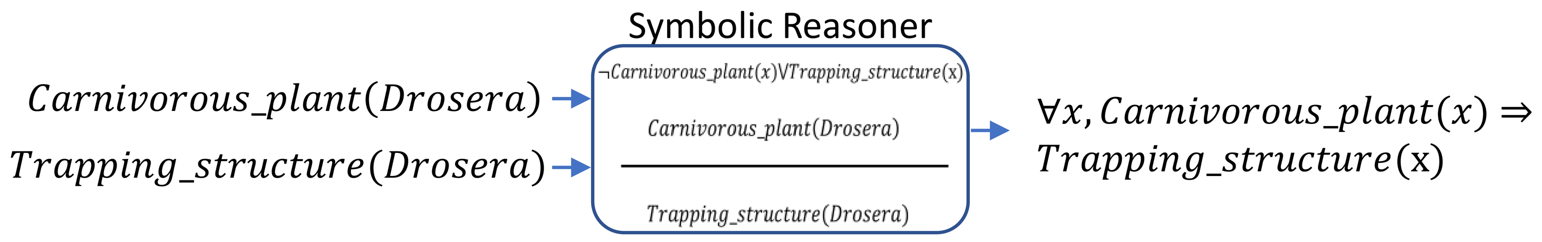
Zonglin Yang [1] Li Dong [2] Xinya Du [3] Hao Cheng [2] Erik Cambria [1] Xiaodong Liu [2] Jianfeng Gao [2] Furu Wei [2]

[1]Nanyang Technological University [2]Microsoft Research [3]University of Texas at Dallas

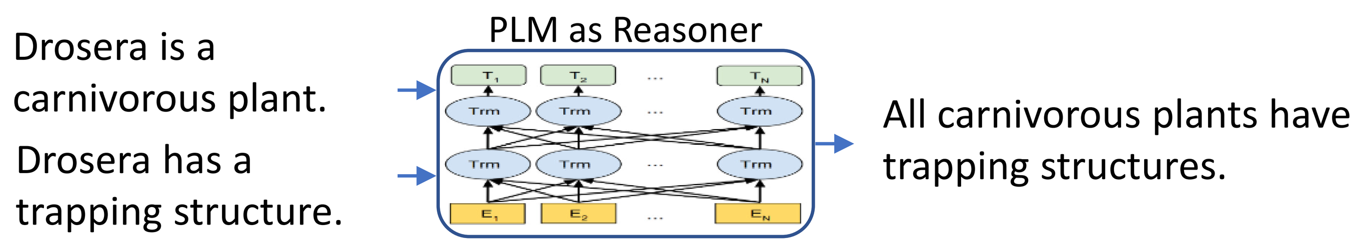
Highlights

- The first work on generative inductive reasoning — in the sense of deriving explicit natural language hypotheses from observations.
- Connected to the classic AI literature, which is the previous computational paradigm of generative inductive reasoning. Advantages are compared.
- Connected to the philosophy literature, which handles the definition of inductive reasoning. Based on the them, we identify key requirements for inductive reasoning from philosophy literature.
- A new dataset for generative inductive reasoning.
- A method based on the key requirements, with a Bayesian design.
- A comprehensive analysis on how LLMs performs on generative inductive reasoning.

A New Paradigm for (Generative) Inductive Reasoning



(a) Formal language as knowledge representation and symbolic reasoner



(b) Natural language as knowledge representation and PLM as reasoner

Systematic Disadvantages of the Previous (Classic AI) Paradigm

Past research works on (generative) inductive reasoning within computer science are investigated by Inductive Logic Programming (ILP), which adopts a classic AI paradigm.

It has three systematic disadvantages:

1. Heavily relying on human effort to transform raw inputs such as natural language and images into symbolic declarative form.
2. Very sensitive to label error.
3. Have no semantic understanding of symbols, resulting in low utilization of annotated data.

The new paradigm can nearly perfectly deal with these systematic issues!

Key Requirements for Inductive Reasoning's Hypotheses

The definition and requirements of inductive reasoning are handled by philosophy research [?].

There are three key requirements:

1. The hypothesis should be deductively consistent with the observations.
2. The hypothesis should reflect the reality.
3. The hypothesis should generalize (covers a larger information scope) than the observations.

We additionally add a requirement in the NLP context:

4. The hypothesis should be clear and meaningful.

Dataset Construction

We construct a dataset (named DEER).

- DEER is to analyze LLMs' generative inductive reason ability.
- DEER is fully constructed by an expert (an author of this paper).
- DEER consists of 200 hypotheses, where each hypothesis is annotated with 3 short observations and 3 long observations.
- DEER adopts a relatively open-ended generation, rather than fixed options.

We focus on rules with the following template:

Rule Template (First Order Logic)	Rule Template (Natural Language)
$\forall x, condition(x) \implies conclusion$	If __, then __.
$\exists x, condition(x) \implies conclusion$	There exists __, which __.
$\forall x, condition(x)\ [\wedge\ condition(x)]^{+} \implies conclusion$	If __ and __, then __.
$\forall x, condition(x)\ [\vee\ condition(x)]^{+} \implies conclusion$	If __ or __, then __.

Table 1. The mapping relation between basic first-order logic rule template and natural language rule template.

An example from DEER:

Short fact 1	Short fact 2	Short fact 3	Rule
The Venus flytrap is a **carnivorous plant** native to... It catches its prey-chiefly insects and arachnids—with a **trapping structure**...	Pitcher plants are several different **carnivorous plants** which have modified leaves known as **pitfall traps**...	Drosera...is one of the largest genera of **carnivorous plants**... The **trapping** and digestion mechanism of Drosera usually employs...	If a plant is **carnivorous**, then it probably has a **trapping** structure.

Analysis Regarding Various Input

Models	Long facts 1 full facts	Short facts 1 full facts	Short facts 2 full facts	Short facts 3 full facts	Short facts 3 missing facts
R+F	9.35 / 2.16	10.87 / 2.33	11.16 / 2.36	11.20 / 2.37	11.52 / 2.40
M1	23.12 / 3.40	24.75 / 3.52	25.22 / 3.55	25.28 / 3.56	24.67 / 3.51
M1+M2	23.43 / 3.49	25.30 / 3.68	25.88 / 3.74	25.68 / 3.68	25.01 / 3.58
M1+M3	23.25 / 3.44	24.91 / 3.55	25.32 / 3.57	25.39 / 3.57	24.77 / 3.52
M1+M4	23.65 / 3.52	25.48 / 3.65	26.04 / 3.73	26.12 / 3.74	25.09 / 3.59
M1+M5	23.23 / 3.44	24.81 / 3.54	25.31 / 3.58	25.28 / 3.55	24.81 / 3.57
CoLM	**24.03** / **3.60**	**25.89** / **3.73**	**26.71** / **3.85**	**26.44** / **3.78**	**25.41** / **3.65**

Table 2. Analysis of PLM (GPT-J)'s performance (measured in METEOR / GREEN) with different input lengths and whether fact contains enough information.

Error Analysis

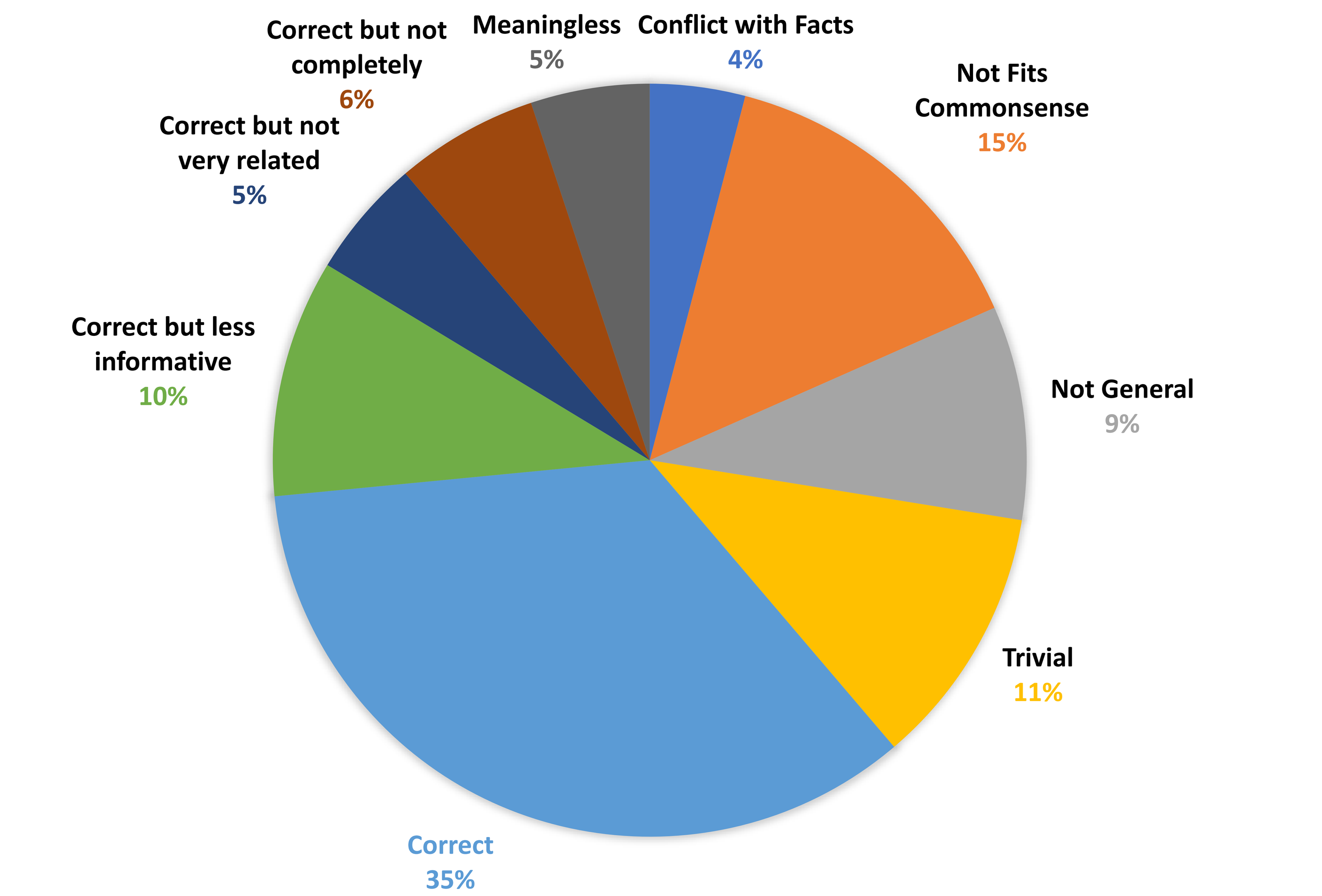


Figure 1. Error Analysis of CoLM with finetuned Module 2/3/4/5. In total 100 rules are manually checked.

Methods Inspired from Key Requirements

We denote $P(A)$ as the probability indicating whether A is valid for simplicity. The framework can be described in a Bayesian design. Specifically, we can denote $P(fact|rule)$ as $P_{M2}(fact|rule)P_{M4}(fact|rule)$, and denote $P(rule)$ as $P_{M3}(rule)P_{M5}(rule)$.

Therefore, the full $P(rule|fact)$ can be approximated as:

$$P(rule|fact) \approx P(fact|rule)P(rule) \approx P_{M2}(fact|rule)P_{M3}(rule)P_{M4}(fact|rule)P_{M5}(rule) \tag{1}$$