

ChatLogic: Integrating Logic Programming with Large Language Models for Multi-Step Reasoning

Jiamou Liu

University of Auckland

Neuro-Symbolic Learning and Reasoning in the Era of Large Language Models @ AAI 2024



THE UNIVERSITY OF
AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

<https://openreview.net/forum?id=AOqGF7Po7Z>

About Ourselves



Jiamou LIU

- Associate Professor at the School of Computing at the University of Auckland
- Leader of LIU AI LAB, an artificial intelligence research and industrial laboratory
- Lead 15+ doctoral students to work on Web3, AI Agent, and multi-modal research



Zhongsheng WANG

- Master student of Data Science at the University of Auckland
- Team member at LIU AI LAB
- Interested in LLM-Based Agent and Natural Language Processing

Content

- Research Background
- Example of Multi-Step Reasoning
- ChatLogic Overview
- Demo for ChatGPT vs. ChatLogic(ChatGPT)
- Result

Research Background

Propositions:

$E \rightarrow H$ ①
 $A \rightarrow B$ ②
 $C \rightarrow E$ ③
 $A \rightarrow D$ ④
 $B \rightarrow C$ ⑤
 $A \rightarrow C$ ⑥
 $E \rightarrow F$ ⑦

How can LLM determine:
 $A \rightarrow F$ True Or False?

There exists a multi-step reasoning path ②⑤③⑦:
 $A \rightarrow B \rightarrow C \rightarrow E \rightarrow F$

- The reasoning ability of LLMs gradually increases with the increase of model parameters
- Existing LLMs have limited ability to extract effective information from long and scrambled texts, especially in **multi-step reasoning questions**
- One possible way is to let LLMs learn reliable code generation rules, and the generated code is used to perform the results of multi-step reasoning questions

Example of Multi-Step Reasoning

Facts:

The bald eagle is sleepy.

The bald eagle is rough.

The leopard is heavy.

The bald eagle visits the rabbit.

The leopard sees the dog.

The rabbit is nice.

The dog is nice.

Rules:

If something is not round then it is heavy.

If something is not strong then it is cute.

If something is furry then it is lovely.

If something is lovely and not big then it is small.

If something is heavy and not round then it is awful.

Query:

The bald eagle is awful.

Result: true

Close-world Assumption:

① The bald eagle is **not round**. (not mentioned in data)

② If something is not round then it is heavy.

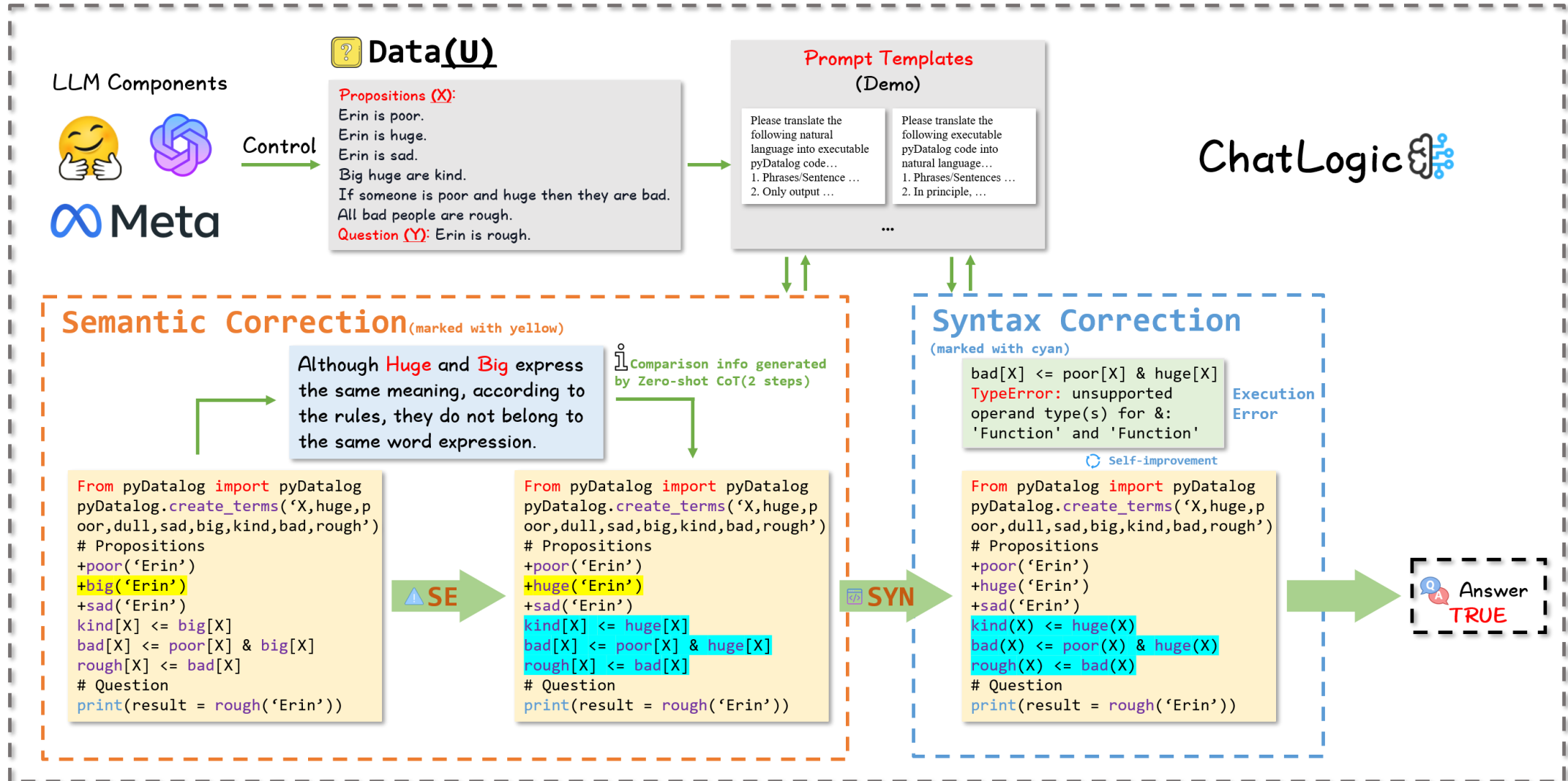
③ If something is heavy and not round then it is awful.

From ① and ②, we know: ④ The bald eagle is heavy.

From ④ and ③, we know: ⑤ The bald eagle is awful.

⑤ is actually the query, the answer is: **True**.

ChatLogic Overview



ChatGPT vs. ChatLogic(ChatGPT)

Propositions:

Facts:

Anne is huge. Anne is strong. Fiona is thin.
Fiona is little. Alan is clever. Dave is rough. Dave is bad.

Rules:

If someone is not big then they are rough.
(Continue on the right)

ChatGPT:

To determine whether "Anne is tiny" is correct, let's examine the given propositions step-by-step and follow the chain of reasoning.

From: Anne is huge.

Based on Proposition 8: "If someone is not big then they are rough. **“Since Anne is huge, she is big.”** Hence, she is not rough.

We need to determine if Anne can be categorized as small and tiny.

(WRONG! "Huge" is not the same as "big". We should not infer downward based on "Anne is big", but based on the closed-world assumption: The proposition: “Anne is big” is not clearly defined, we should think that Anne is not big.)

Therefore, we don't have enough evidence to say that "Anne is tiny" based on the given propositions.

Answer: 0 (The statement "Anne is tiny" is wrong based on the provided propositions.)

If someone is not sad then they are kind.

If someone is clever then they are quiet.

If someone is quiet and not poor then they are smart.

If someone is rough and not big then they are dull.

If someone is dull then they are small.

Question:

Anne is tiny?

ChatLogic:

```
# Declare the pyDatalog variables
pyDatalog.create_terms('X,huge,strong,little,
rough,clever,sad,kind,quiet,tiny,big [...]')
# Define the facts
+huge('Anne')
+strong('Anne')
+little('Fiona') [...]
# Define the rules
rough(X) <= ~big(X)
kind(X) <= ~sad(X)
quiet(X) <= clever(X) [...]
# Query the knowledge base
result = tiny('Anne')
if result:
    print(1)
else:
    print(0)
```

result: 1
Anne is tiny.

Result on 3 Datasets

PARARULE-Plus¹

Model	Method	Depth=2	Depth=3	Depth=4	Depth=5	Total
GPT-3.5	Base	0.4	0.34	0.32	0.3	0.344
	Zero-shot CoT	0.42	0.42	0.41	0.3	0.3875
	ChatLogic	0.49	0.56	0.65	0.41	0.5275
GPT-4	Base	0.65	0.75	0.42	0.4	0.555
	Zero-shot CoT	0.72	0.72	0.62	0.7	0.69
	ChatLogic	0.72	0.8	0.7	0.7	0.73
Llama 2-7B	Base	0.11	0.06	0.01	0.01	0.0475
	Zero-shot CoT	0.15	0.13	0.08	0.06	0.105
	ChatLogic	0.2	0.13	0.22	0.18	0.1825

Test results of 2 modules

Dataset	Model	Base	SE	SE+SYN
CONCEPTRULES V1	GPT-3.5	0.63	0.68	0.7
	GPT-4	0.92	0.96	0.96
	Llama 2-7B	0.31	0.60	0.62
CONCEPTRULES V2	GPT-3.5	0.6	0.73	0.8
	GPT-4	0.92	0.93	0.95
	Llama 2-7B	0.33	0.52	0.53
PARARULE-Plus	GPT-3.5	0.26	0.5	0.62
	GPT-4	0.54	0.64	0.7
	Llama 2-7B	0.1	0.16	0.16

SE = Semantic Correction

SYN = Syntax Correction

CONCEPTRULES V1² & V2³

Model	Method	CONCEPTRULES V1		CONCEPTRULES V2	
		simplified	full	simplified	full
GPT-3.5	Base	0.57	0.55	0.5	0.51
	Zero-shot CoT	0.63	0.51	0.7	0.67
	ChatLogic	0.69	0.67	0.79	0.74
GPT-4	Base	0.95	0.94	0.89	0.86
	Zero-shot CoT	0.96	0.97	0.95	0.94
	ChatLogic	0.96	0.96	0.94	0.94
Llama 2-7B	Base	0.32	0.29	0.31	0.24
	Zero-shot CoT	0.42	0.41	0.33	0.3
	ChatLogic	0.48	0.49	0.37	0.36

1 <https://github.com/Strong-AI-Lab/PARARULE-Plus>

2 <https://bit.ly/3uVemXG>

3 <https://bit.ly/3PApIIB>