

ARE YOU TALKING LOGIC TO ME? ASSESSING LANGUAGE MODELS SYLLOGISTIC REASONING CAPABILITIES



Hanna Abi Akl, Fabien Gandon, Catherine Faron, Pierre Monnin

RuleML+RR 2026



Content

Motivation

An Invitation to Logic

Methodology

Experiments

Findings

Take-Home Message

CLGC in the Wild



Motivation

- ▶ Language Models (LMs) exhibit potential in logical reasoning tasks
- ▶ A logical reasoning problem can be formalized in different ways
- ▶ We revisit LM reasoning on syllogisms
 - How well can LMs reason on syllogisms?
 - What happens if we formalize the syllogism differently?
- ▶ Research Questions (RQs)
 - RQ1. How do different input formal notations impact LM syllogistic reasoning?
 - RQ2. How does prompting LMs with pre-defined syllogism categories affect their reasoning?



An Invitation to Logic

- ▶ Syllogisms = using 2 or more statements to derive a new one $\rightarrow$ extending knowledge
- ▶ Logical Notations = Knowledge Representation (KR) languages to represent logical statements
- ▶ Examples: CLIF & CGIF (Sowa, 1992), TFLPLUS (Purdy et al., 2002), CLINGO (Gebser et al., 2014), MINIFOLx (ours)

Table 1. Notation equivalent for a Natural Language (NL) statement.

Notation	Example
NL	If someone has the flu, then they have influenza
FOL	$\forall x (\text{Has}(x, \text{flu}) \rightarrow \text{Has}(x, \text{influenza}))$
CLIF	forall x (has(x, flu) implies has(x, influenza))
CLINGO	forall (has(x, flu) :- has(x, influenza))
CGIF	@every *x [(has[(?x flu)] has[(?x influenza)])]
MINIFOL2	all:x (has(x, flu) :- has(x, influenza))
TFLPLUS	$-(+H0-+H0)$

Methodology

▶ **Datasets.** FOLIO & P-FOLIO

- Set of premises and conclusions in Natural Language (NL) and First-Order Logic (FOL) notations

▶ **Experiment.** Predict if conclusions are **True**, **False** or **Unknown** from premises

- Extended experiments from FOLIO paper ((Han et al., 2024)) in *SFT and **ZS

▶ **Framework.** Syllogistic Evaluation Framework-Common Logic Grammar Construction

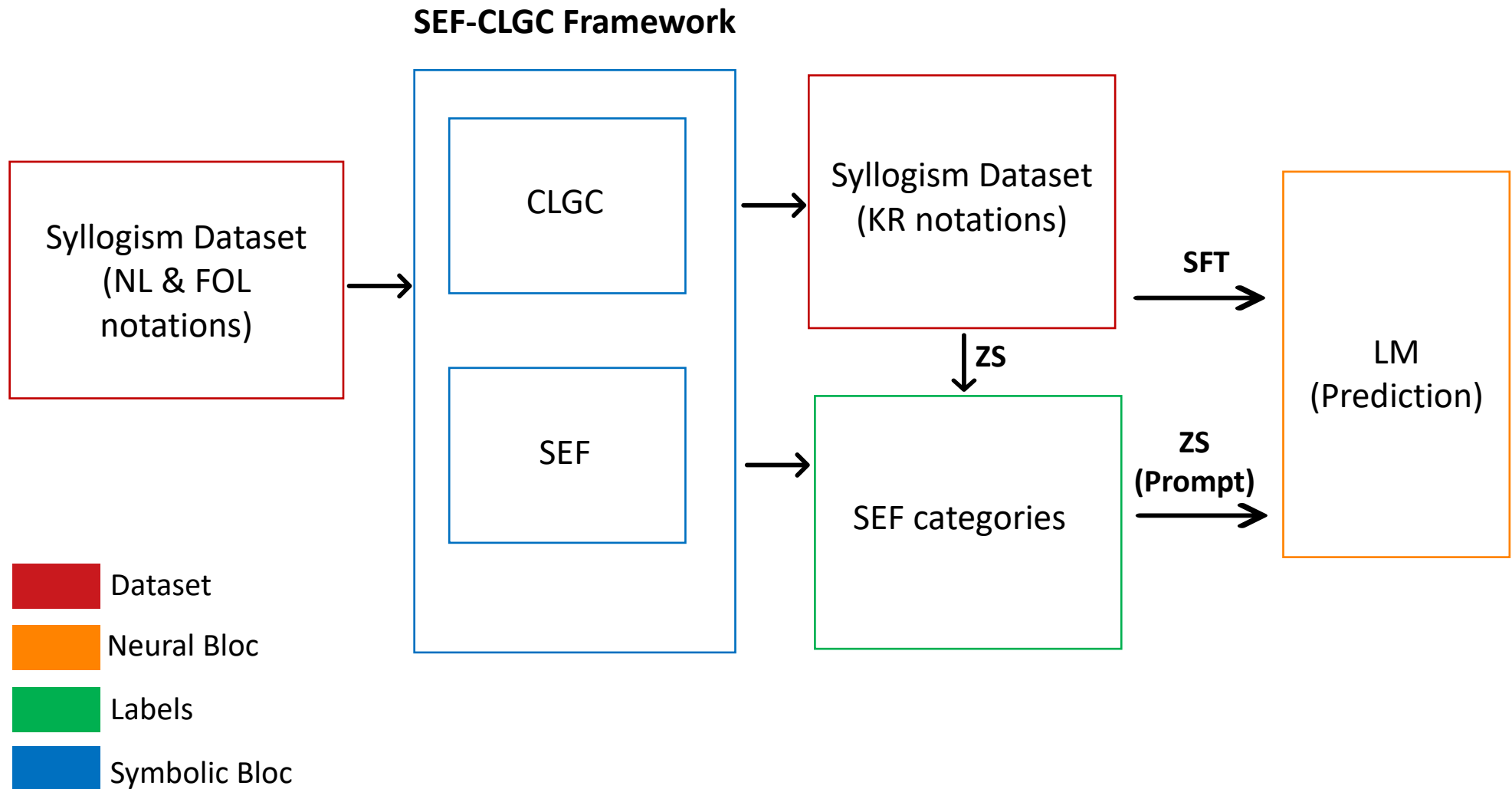
▶ **Models.** Focus on small language models

- SFT: Flan-T5-small, Flan-T5-large
- ZS: Gemma-2-2b-it, Llama-3.2-3b-it, Phi-3.5-mini-it

**Supervised Fine-Tuning*

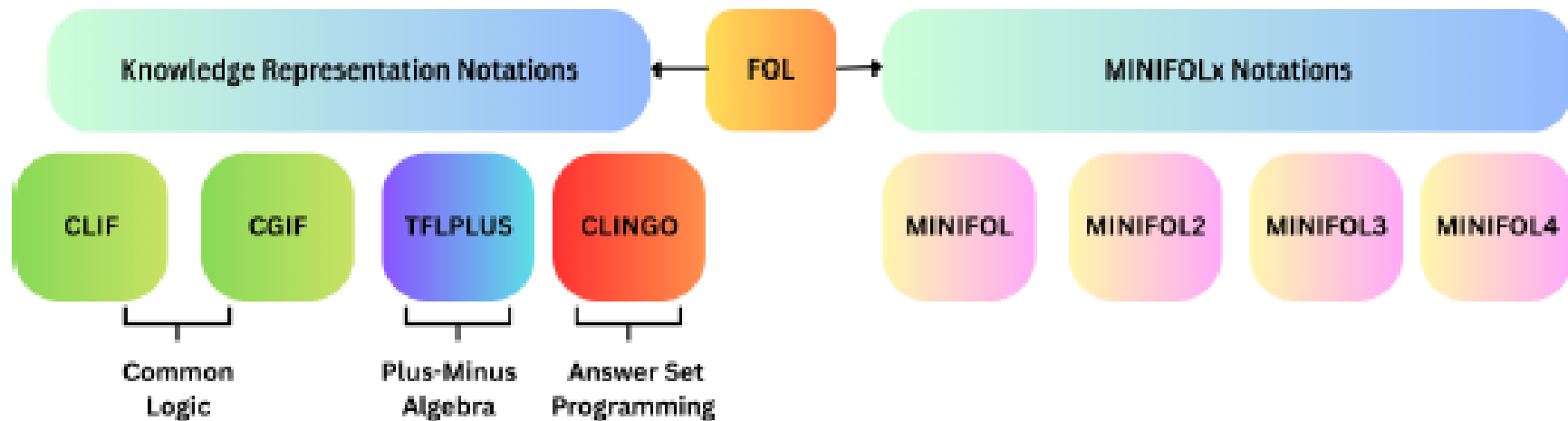
***Zero-Shot prompting*

Methodology



Methodology - CLGC

- ▶ CLGC = Logical notation generation from FOL
- ▶ Notations are chosen based on pre-defined criteria (e.g. verbosity, frequency)



Methodology - SEF

- ▶ SEF = Categorize syllogisms into types of reasoning (Wu et al., 2023)
- ▶ Each category indicates a type/structure of syllogism (and by extension, reasoning)

SEF Category	Syllogism example
Complex	RenamedAs(fortCarillon, fortTiconderoga) Built(pierredeRigauddeVaudreuil, fortCarillon) LocatedIn(fortCarillon, newFrance)
	$\neg$ LocatedIn(newFrance, europe)
Categorical	SportingEvent(olympics) LastSummerOlympics(tokyo)
	MostMedals(unitedStates, tokyo)
Hypothetical	$\forall x$ (Square(x) $\rightarrow$ FourSided(x)) $\forall x$ (FourSided(x) $\rightarrow$ Shape(x))
	$\forall x$ (Square(x) $\rightarrow$ Shape(x))
Disjunctive	FlyTo(susan, lgaAirport) $\forall x \forall y$ (FlyFrom(x, y) $\oplus$ FlyTo(x, y))
	FlyFrom(john, lgaAirport)

Experiments

► SFT (Train/Val/Test splits)

- Training (Train & Val splits): Syllogism in KR notation + Label
- Evaluation (Test split): Syllogism in KR notation → Model predicts Label

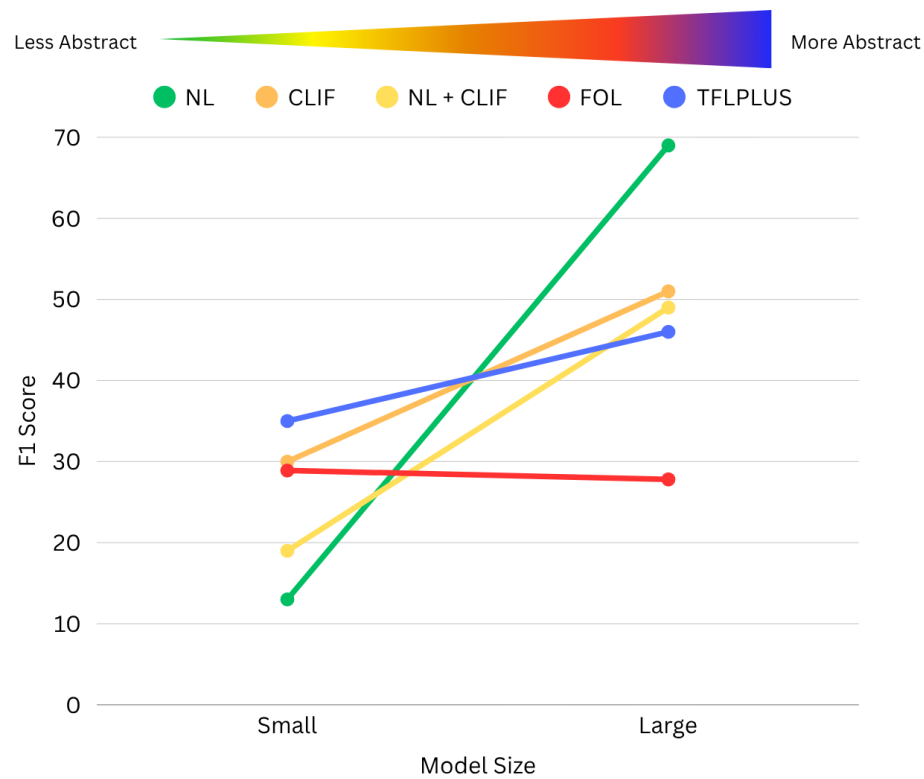
► ZS (2 scenarios)

- Scenario 1: Prompt containing syllogism in KR notation → Model predicts Label
- Scenario 2: Prompt containing syllogism in KR notation + SEF category → Model predicts Label

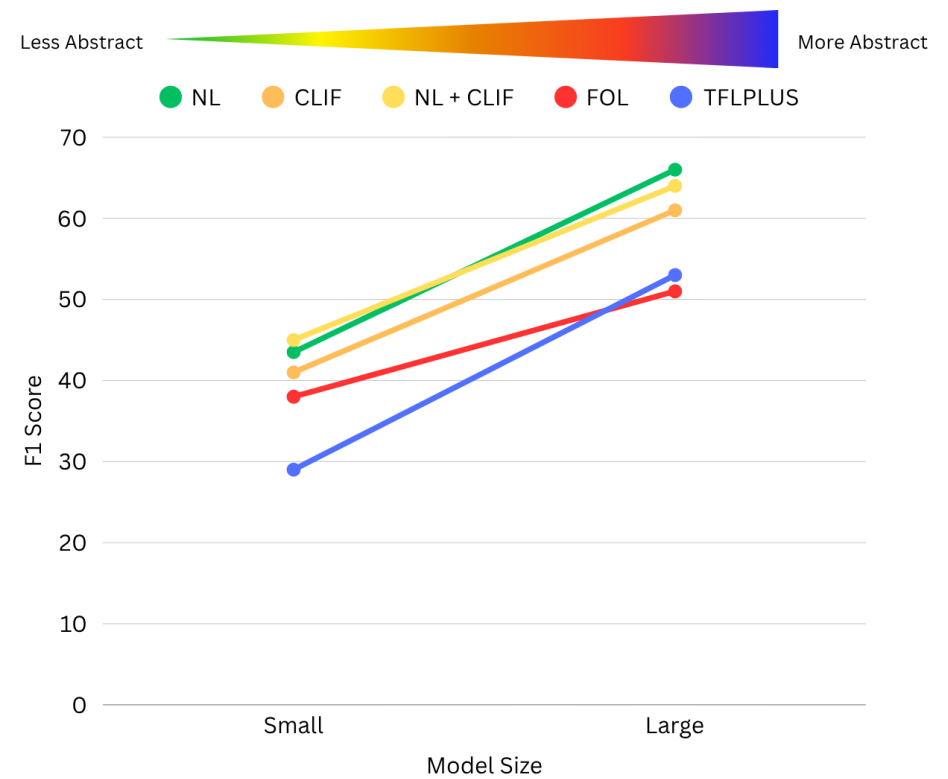
Dataset	Total	Label			SEF			
		True	False	Unknown	Categorical	Hypothetical	Disjunctive	Complex
FOLIO-KR	1204	460	351	393	17	67	653	467
P-FOLIO-KR	301	122	69	110	9	13	40	239

Findings

► Notation performance and model size



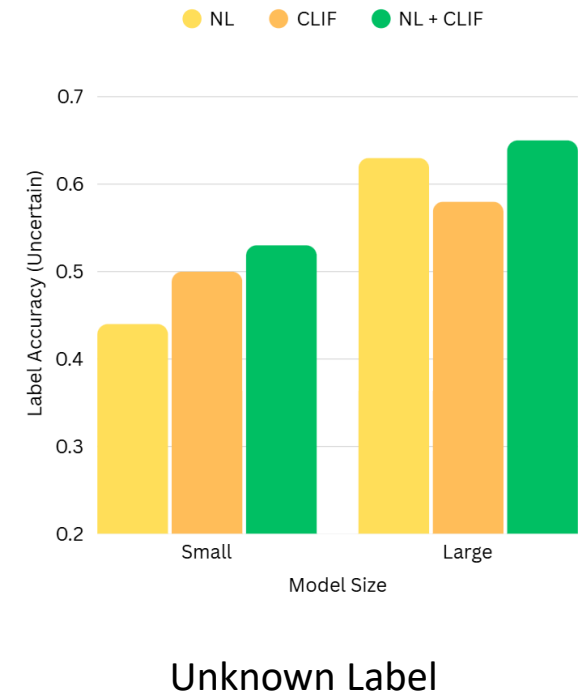
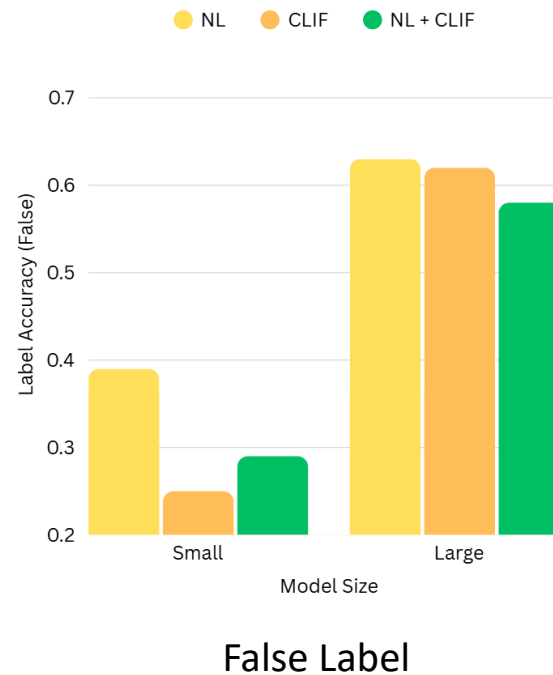
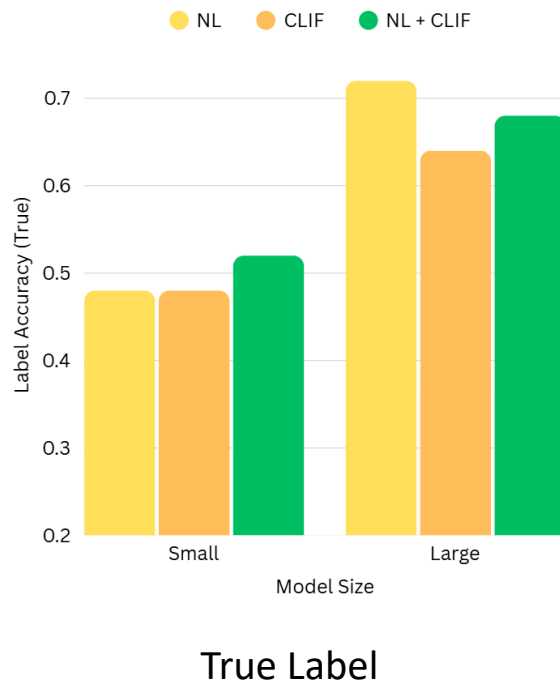
P-FOLIO-KR



FOLIO-KR

Findings

► Notation and type of reasoning



Findings

- ▶ Performance increases with model size independently of abstraction (**SFT**)
- ▶ Notation performance depends on model and dataset (**SFT**)
- ▶ Combining notations yields more conservative reasoning (**SFT**)
- ▶ NL + CLIF as a conservative reasoning language representation (**SFT**)
- ▶ Choice of notation depends on model and training paradigm (**ZS**)
- ▶ SEF category impact varies with model and notation (**ZS**)
- ▶ Logical notations show faster reasoning inference than NL (**ZS**)



Take-Home Message

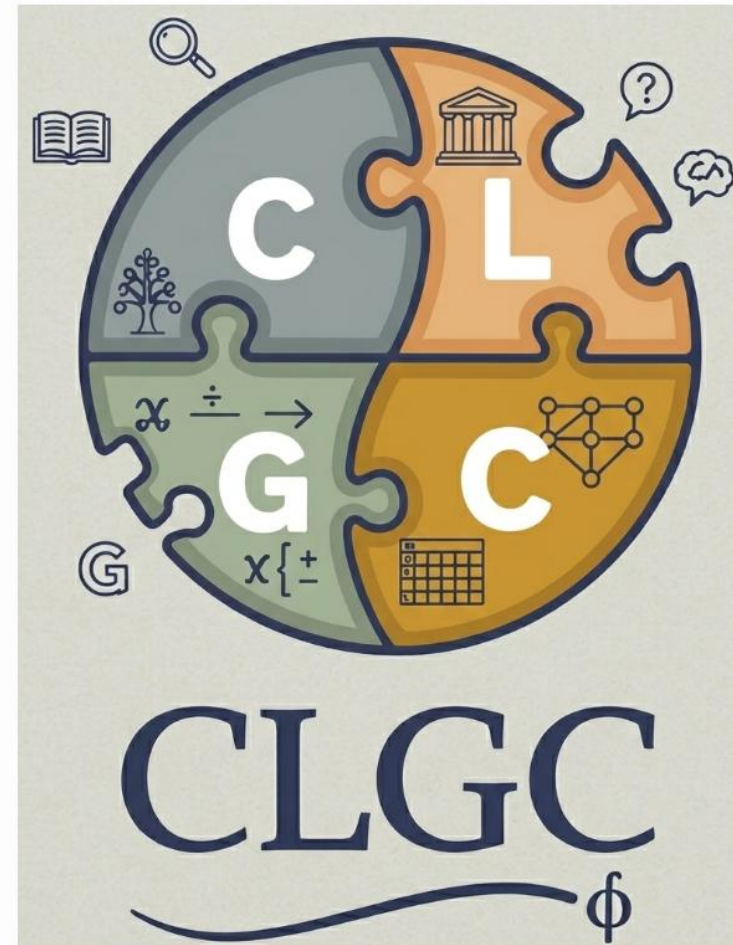
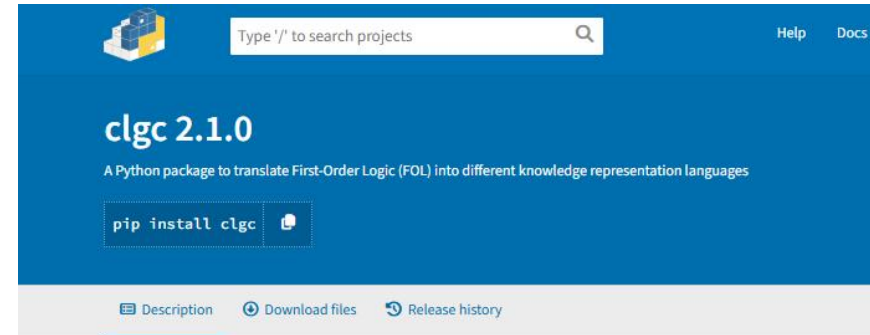
Different KR notations affect LM reasoning on logical problems

Multi-stage neuro-symbolic KR pipeline to steer logical reasoning



CLGC in the Wild

- ▶ Open-source package
- ▶ Automatic generation of KR notations
- ▶ Supports syllogism datasets
- ▶ SEF labeling



Thank You.

