Coupling Formal Logic and Natural Language Reasoning

Jinu Lee (UIUC)

02/28

NRF-BRL (Human-AI Collaborative Programming)

About me

Jinu Lee

Education

Ph.D.	University of Illinois Urbana-Champaign *(Advised by Dr. Julia Hockenmaier)*	*Aug 2024 – Now*
B.S.	Seoul National University	*Mar 2018 - Aug 2024*

Work experience

Research Intern	Microsoft Research	*May 2025 - Aug 2025*
Research Engineer	LBOX	*Jul 2023 - Jun 2024*
Research Intern	NCSOFT Language AI Lab	*Jun 2020 - Nov 2020*

Contents

Formal logic and language

"There is no important theoretical difference between the natural and the artificial languages."

\- R. Montague (1974)

Formal logic and language

Natural language and formal logic are analogous in **reasoning**

※ *Deductive Reasoning: Do the premises **entail** hypothesis?*

p1. Socrates is human.

p2. All humans die.

Human language instincts

h. Socrates will die.

Do *p*'s entail *h*?

p1. human(Socrates)

p2. $\forall$x(human(x) $\rightarrow$ die(x))

Formal inference rules
e.g. Natural deduction

h. die(Socrates)

Formal logic and language

Natural language and formal logic have strengths/weaknesses:

Natural language	Formal logic
• Express diverse semantics • Fuzzy and ambiguous	• Computable and verifiable • Rigid and brittle

$\rightarrow$ Naturally, we should try **using both** to complement each other!

Formal logic and language

Q. How to use formal logic representations **for natural language reasoning**?

A. **Parse-then-execute** pipeline:

(1) Translate NL to logic → (2) Execute automatic provers

e.g. Given that (p1) Socrates is human and (p2) all humans die, (h) will Socrates die?

p1. Socrates is human.

p2. All humans die.

h. Socrates will die.

Olausson et al., (2023) LINC
Pan et al., (2023) Logic-LM
Han et al., (2024) FOLIO
inter alia.

Formal logic and language

Q. How to use formal logic representations **for reasoning**?

A. **Parse-then-execute** pipeline:

(1) Translate NL to logic → (2) Execute automatic provers

e.g. Given that (p1) Socrates is human and (p2) all humans die, (h) will Socrates die?

1. Parse

p1. Socrates is human.	p1. human(Socrates)
p2. All humans die.	p2. $\forall$x(human(x) $\rightarrow$ die(x))
h. Socrates will die.	h. die(Socrates)

Olausson et al., (2023) LINC
Pan et al., (2023) Logic-LM
Han et al., (2024) FOLIO
inter alia.

Formal logic and language

Q. How to use formal logic representations **for reasoning**?

A. **Parse-then-execute** pipeline:

(1) Translate NL to logic → (2) Execute automatic provers

e.g. Given that (p1) Socrates is human and (p2) all humans die, (h) will Socrates die?

1. Parse

p1. Socrates is human. → p1. human(Socrates)

p2. All humans die. → p2. $\forall$x(human(x) $\rightarrow$ die(x))

h. Socrates will die. → h. die(Socrates)

2. Execute

Automatic theorem provers → **True** (h is deduced by p's)

Prover9

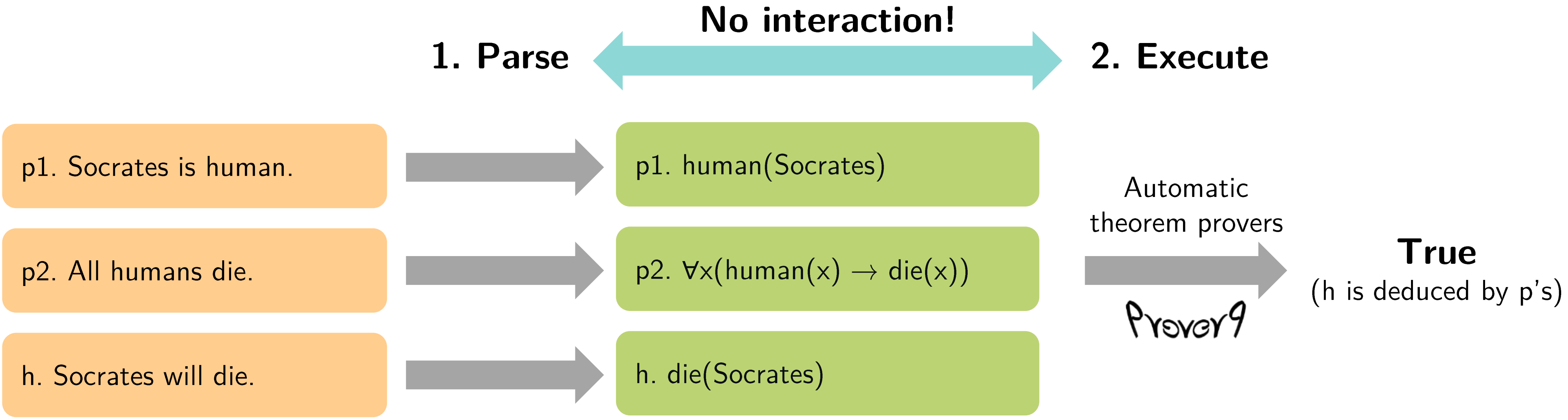
Olausson et al., (2023) LINC
Pan et al., (2023) Logic-LM
Han et al., (2024) FOLIO
inter alia.

Formal logic and language

In parse-then-execute, semantic parsing and reasoning are **decoupled**

- Semantic parsers are not aware of following reasoning process
- Provers blindly rely on the semantic parses

→ How to model **the interaction** between semantic parsing and reasoning?



Formal logic and language

Key question: How to model the **interaction** between semantic parsing and execution?

- **Interleaving** semantic parsing and execution
  - *Work 1: Symbolic Backward Chaining*
- Using desired execution results as **training objective** for parsers
  - *Work 2: Entailment-preserving FOL representations*

Contents

Motivation

How to solve *complex* reasoning problems with lots of premises?

- Chain-of-thoughts (step-by-step reasoning) is the de facto standard
- Still, there are alternative approaches: *e.g.* **Backward chaining**

Problem

The battery charge in Mary's cordless vacuum cleaner lasts ten minutes. It takes her four minutes to vacuum each room in her house. Mary has three bedrooms, a kitchen, and a living room. How many times does Mary need to charge her vacuum cleaner to vacuum her whole house?

Solution

Mary has 3 + 1 + 1 = 5 rooms in her house.
At 4 minutes a room, it will take her 4 * 5 = 20 minutes to vacuum her whole house.
At 10 minutes a charge, she will need to charge her vacuum cleaner 20 / 10 = 2 times to vacuum her whole house.

Final Answer

2

Hunter et al. (2024) GSM8k

Motivation

Backward chaining(=top-down reasoning): Decomposes problems to subproblems (Divide&Conquer)

Algorithmic solution for backward chaining: **SLD resolution** in logic programming (Prolog)

Clark et al., (2021) ProofWriter

Fact 1. is(alan, young). *Alan is young.*
Fact 2. is(bob, young). *Bob is young.*
Fact 3. is(bob, round). *Bob is round.*
Rule 1. is(charlie, cold) :- is(X, young), is(X, round).
If someone is young and round, Charlie is cold.

Goal. is(charlie, cold)? *Is charlie cold?*

Motivation

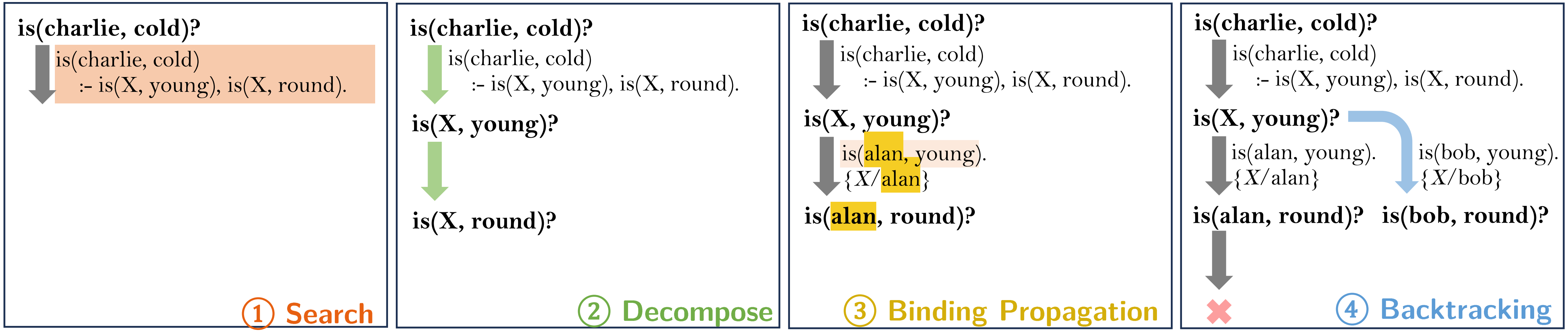
Backward chaining(=top-down reasoning): Decomposes problems to subproblems (Divide&Conquer)

Algorithmic solution for backward chaining: **SLD resolution** in logic programming

Clark et al., (2021) ProofWriter

Fact 1. is(alan, young). *Alan is young.*
Fact 2. is(bob, young). *Bob is young.*
Fact 3. is(bob, round). *Bob is round.*
Rule 1. is(charlie, cold) :- is(X, young), is(X, round).
If someone is young and round, Charlie is cold.

Goal. is(charlie, cold)? *Is charlie cold?*



Motivation

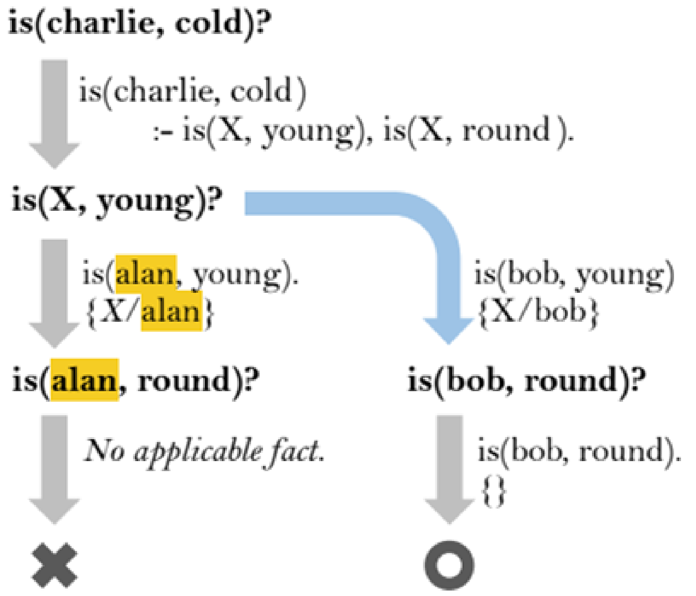
Backward chaining(=top-down reasoning): Decomposes problems to subproblems (Divide&Conquer)

Algorithmic solution for backward chaining: **SLD resolution** in logic programming

Clark et al., (2021) ProofWriter

Fact 1. is(alan, young). *Alan is young.*
Fact 2. is(bob, young). *Bob is young.*
Fact 3. is(bob, round). *Bob is round.*
Rule 1. is(charlie, cold) :- is(X, young), is(X, round).
If someone is young and round, Charlie is cold.

Goal. is(charlie, cold)? *Is charlie cold?*

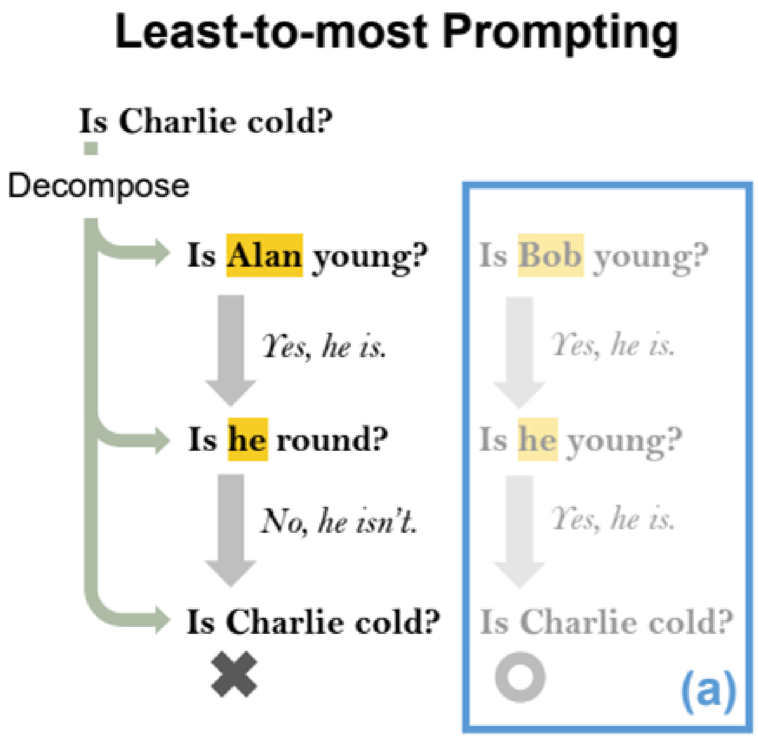


③ Binding Propagation
④ Backtracking

Motivation

Attempts to use LLMs for natural language-based backward chaining:
However, these methods are *incomplete.*

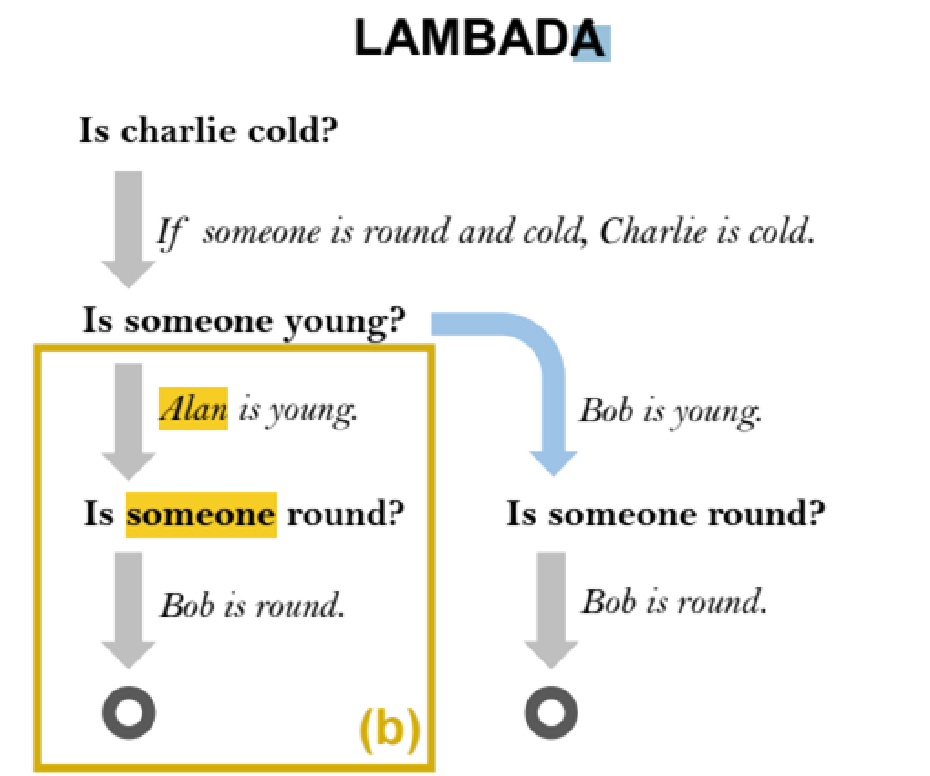
- **Task decomposition** (Least-to-most; Zhou et al., ICLR 2023): No **backtracking**



Motivation

Attempts to use LLMs for natural language-based backward chaining:
However, these methods are *incomplete.*

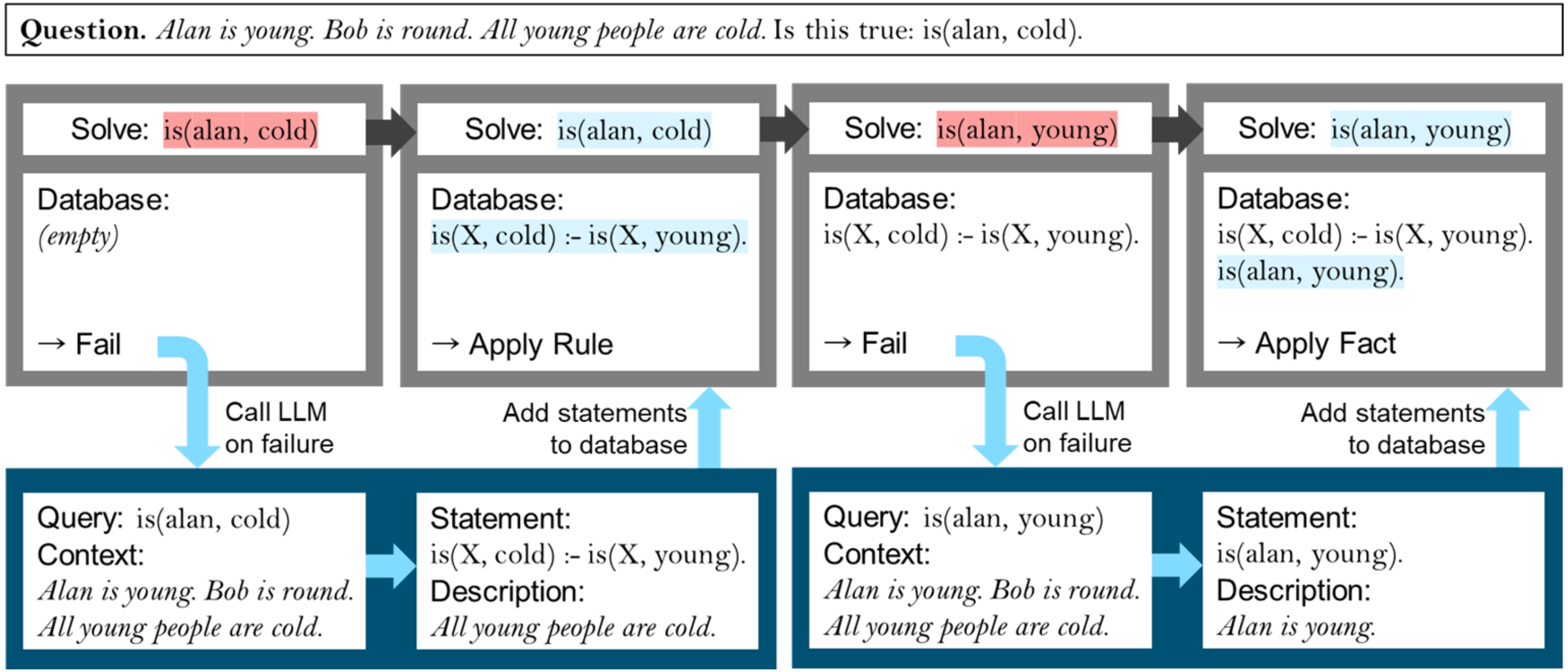
- **Task decomposition** (Least-to-most; Zhou et al., ICLR 2023): No **backtracking**
- **LAMBADA** (Kazemi et al., ACL 2023): No **binding propagation**



Method

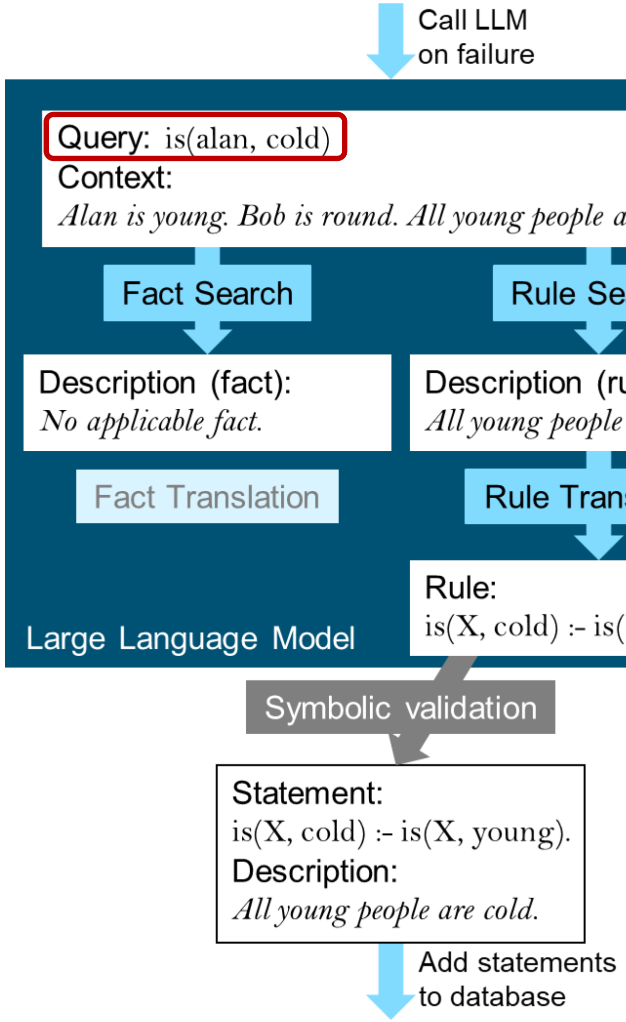
Idea: Interleaving execution (SLD resolution) and semantic parsing (LLM)

1. *Execution*: SLD Resolution solver (gray) searches for the symbolic proof
2. *Semantic parsing:* When reach dead end, ask LLM (navy) to generate rule from input
3. Repeat until a solution is found or no more possible paths are left



Method

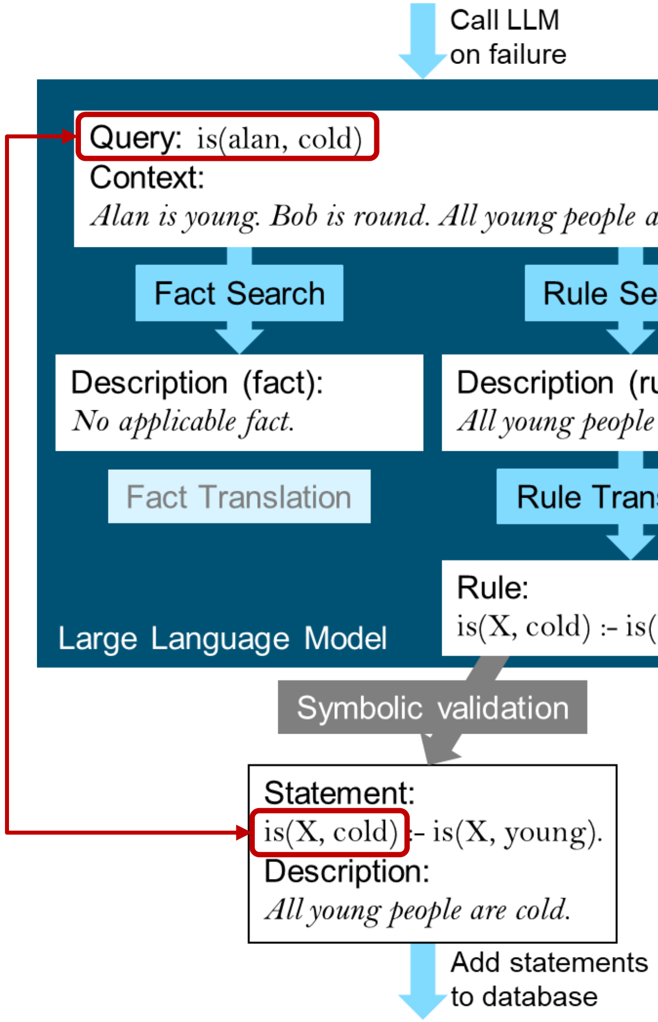
Semantic parsing: LLMs search/generate logic stmts conditioned on the **symbolic query**



Method

Semantic parsing: LLMs search/generate logic stmts conditioned on the **symbolic query**

Symbolic Validation



Experiments

Reasoning performance in 7 benchmarks (deductive, relational, arithmetic)

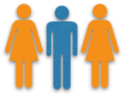
- **Deductive:** If *p1* and *p2*, then *h*
- **Relational:** If *(e1, r1, e2)* and *(e2, r2, e3)*, then *(e1, r3, e3)*
- **Arithmetic:** If $x=N$ and $y=M$, then $z=f(N, M)$

Clark et al., (2021) ProofWriter

Fact 1. is(alan, young). *Alan is young.*
Fact 2. is(bob, young). *Bob is young.*
Fact 3. is(bob, round). *Bob is round.*
Rule 1. is(charlie, cold) :- is(X, young), is(X, round).
If someone is young and round, Charlie is cold.

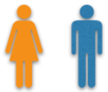
Goal. is(charlie, cold)? *Is charlie cold?*

Kristin and her son **Justin** went to visit her mother **Carol** on a nice Sunday afternoon. They went out for a movie together and had a good time.



Q: How is **Carol** related to **Justin** ?

A: Carol is the **grandmother** of Justin



Sinha et al., (2019) CLUTRR

Problem

The battery charge in Mary's cordless vacuum cleaner lasts ten minutes. It takes her four minutes to vacuum each room in her house. Mary has three bedrooms, a kitchen, and a living room. How many times does Mary need to charge her vacuum cleaner to vacuum her whole house?

Solution

Mary has 3 + 1 + 1 = 5 rooms in her house.
At 4 minutes a room, it will take her 4 * 5 = 20 minutes to vacuum her whole house.
At 10 minutes a charge, she will need to charge her vacuum cleaner 20 / 10 = 2 times to vacuum her whole house.

Final Answer

2

Cobbe et al. (2022) GSM8k

Results

Outperforms backward chaining baselines!

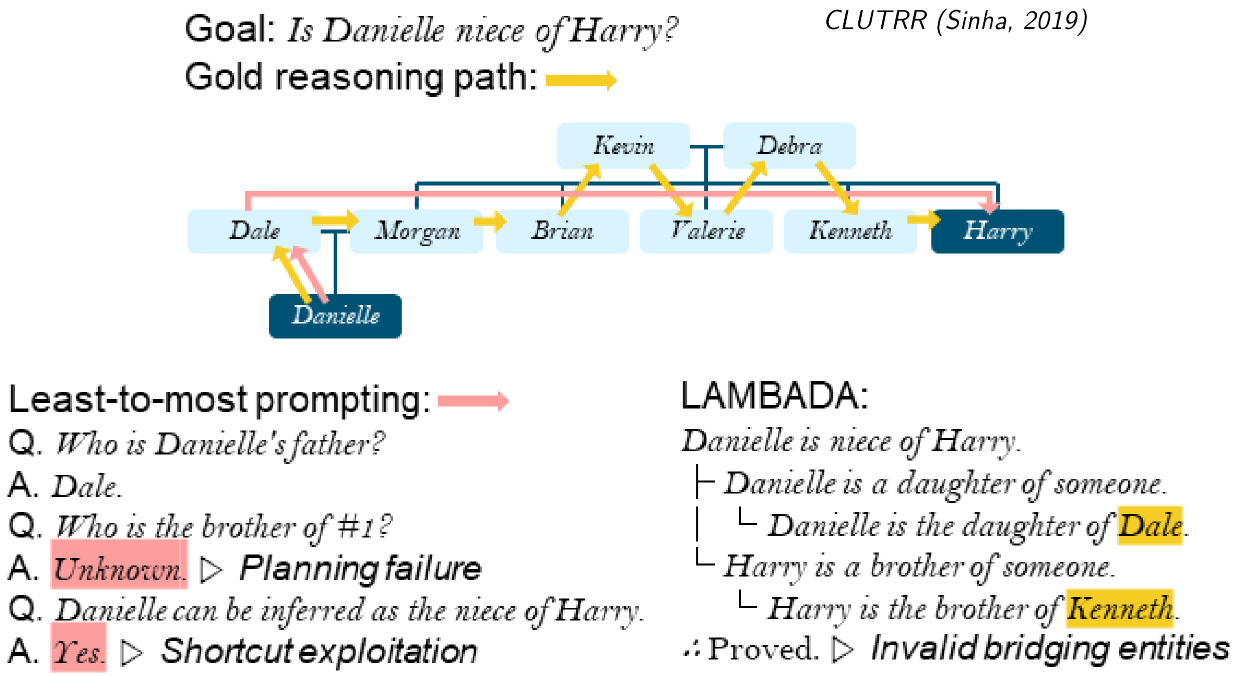
1. Least-to-most(task decomposition) often show shortcut reasoning
2. LAMBADA cannot solve problems that require binding propagation (relational/math)

Model	Method	Deductive				Relational	Arithmetic	
		ProofWriter	BirdsElec	ParaRules	PrOntoQA	CLUTRR	MAWPS	GSM8k
GPT-4	Least-to-most	71.5	88.2	71.8	87.5	81.5	84.3	60.6
	LAMBADA	69.7	83.4	59.7	96.0	73.8	0.0	0.0
	SymBa	**79.8**	**94.4**	**79.2**	**96.3**	**84.3**	**86.7**	**63.8**
Claude-3	Least-to-most	60.3	75.7	54.0	86.0	77.0	**94.2**	59.3
	LAMBADA	69.3	62.7	57.7	67.0	69.0	0.0	0.0
	SymBa	**77.6**	**77.3**	**69.0**	**91.0**	**85.0**	94.1	**67.4**
LLaMa-3	Least-to-most	61.4	71.0	66.7	**95.0**	72.0	**89.0**	61.5
	LAMBADA	64.0	82.3	62.1	90.8	73.3	0.0	0.0
	SymBa	**70.4**	**92.9**	**71.7**	93.3	**90.5**	87.9	**67.0**

Results

Outperforms backward chaining baselines

1. Least-to-most(task decomposition) achieves low proof accuracy
2. LAMBADA cannot solve problems that require binding propagation (relational/math)

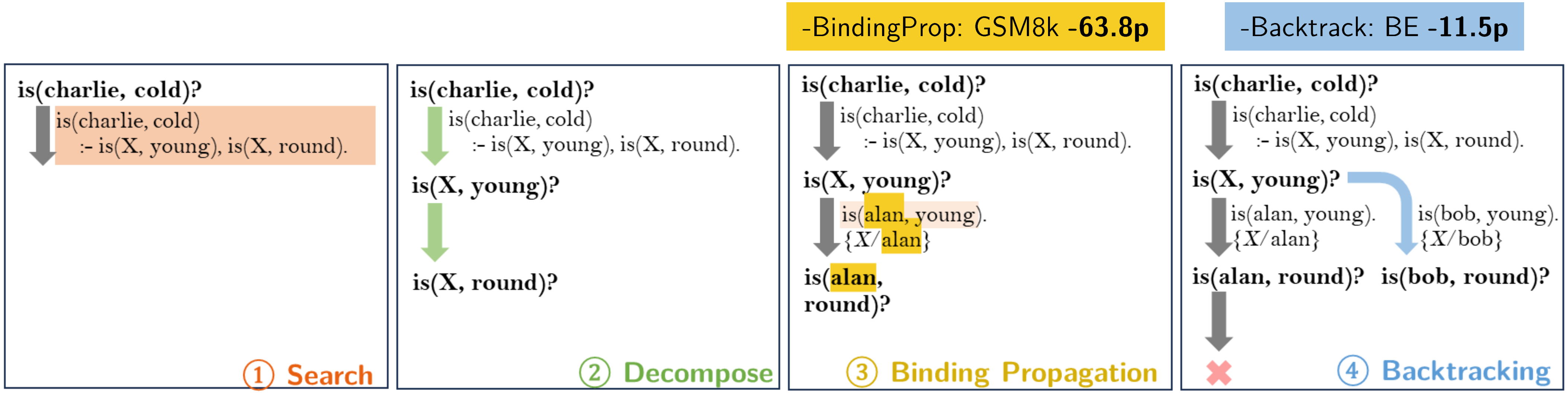


Results

Ablation: removing backtracking and binding propagation from SymBa

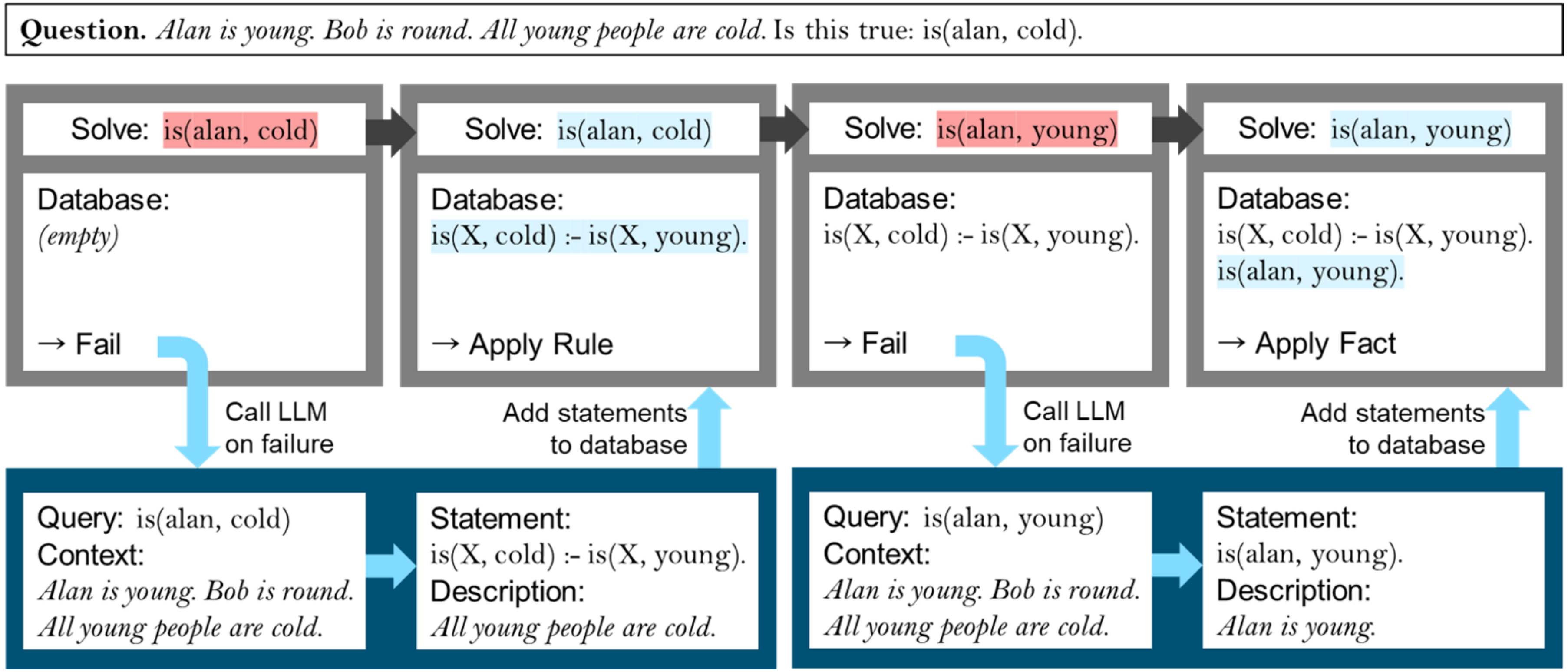
- Backtracking and binding propagation is indeed crucial in performance

	Benchmarks			
	PW	BE	CLUTRR	GSM8k
SymBa	79.8	94.4	84.3	63.8
-Backtrack	76.3	82.9	69.8	62.0
(Least-to-most)	71.5	83.4	81.5	60.6
-BindingProp	80.5	92.2	68.3	0.0
(LAMBADA)	69.7	83.4	73.8	0.0



Conclusion

- Proposed **Symbolic Backward Chaining (SymBa)**
  - *Interleaving* semantic parsing and symbolic inference steps
  - Outperforming language-only backward chaining baselines
  - Showed that *backtracking* and *binding propagation* is crucial for backward chaining



Contents

Introduction

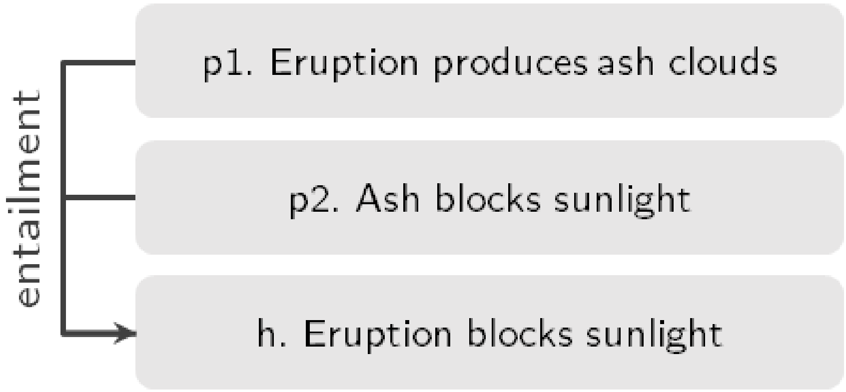
Can formal logic apply to **more natural text**?

i.e. Natural Langauge Entailment (*a.k.a. NLI; RTE*):

- Built from natural texts (non-synthetic)
- **Loose entailment** compared to deductive/relational/arithmetic

"p entails h when a human reading p ***will likely infer that*** *h is also true"*

Dagan et al., (2005). RTE Challenge



Introduction

Expressing natural language into formal logic is ambiguous:

Eruption produces ash clouds $\longrightarrow$

- $\exists x\, \mathrm{eruption}(x) \rightarrow \boxed{\exists y\, \mathrm{ashCloud}(y)}$
- $\exists x\, \mathrm{eruption}(x) \rightarrow \boxed{\exists y\, \mathrm{ash}(y) \land \mathrm{cloud}(y)}$
- $\forall x\, \mathrm{eruption}(x) \rightarrow \exists y\, \mathrm{ash}(y)$

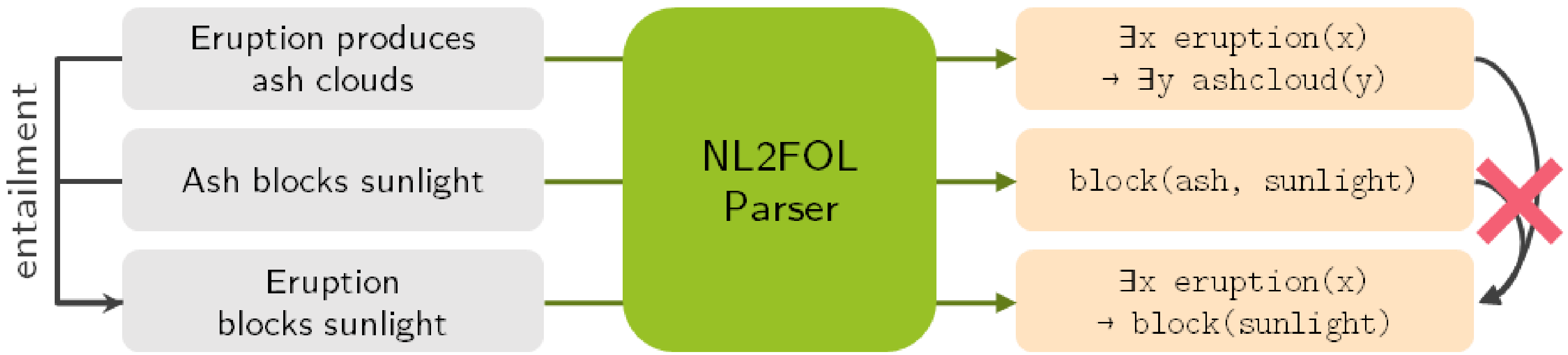
Infamous problems:

- **Arbitrariness**: mapping between NL and predicate is arbitrary
  - $[[$*ash cloud*$]] = \mathrm{ashCloud}(y)$ *vs.* $\mathrm{ash}(y) \land \mathrm{cloud}(y)$
- **Brittleness**: slight different in predicates ban unification
  - $\mathrm{ashCloud}(x) \not\leftrightarrow \mathrm{ash}(x) \land \mathrm{cloud}(x)$

Introduction

Due to such ambiguity, naively translating NL to FOL **might not preserve natural entailment**

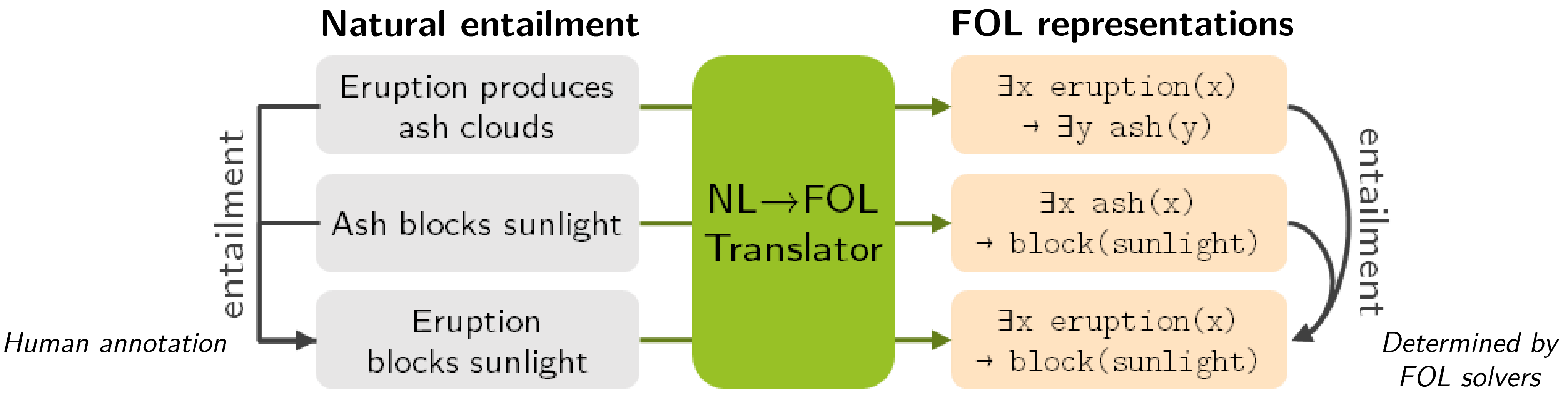
- Serious caveat of *parse-then-execute* approaches



Introduction

Problem:

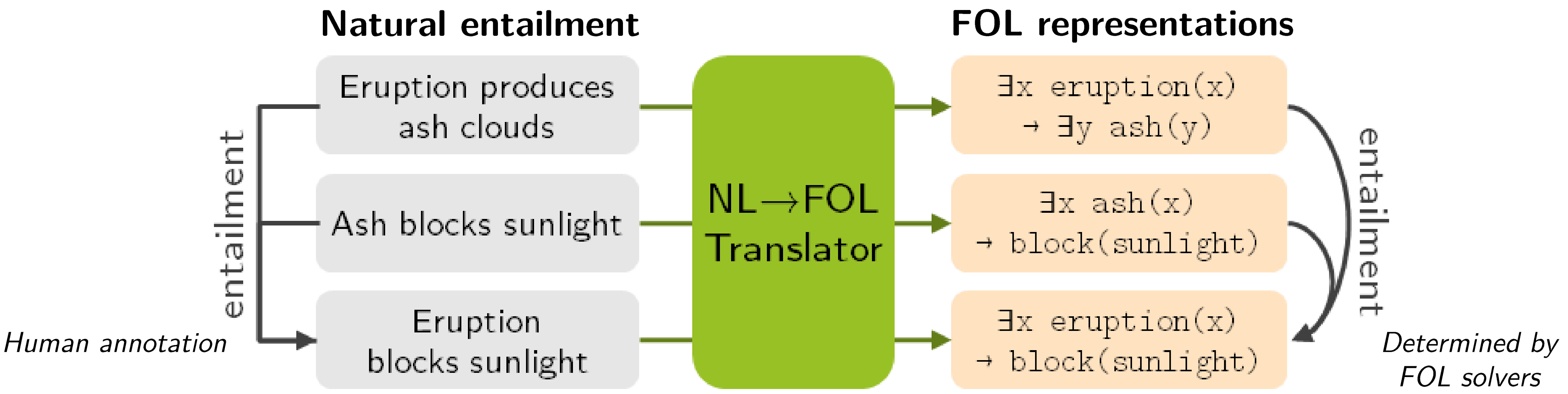
- **First-order logic (FOL)** semantic representations are widely used for logical inference
- However, FOL are limited in expressing **natural entailments**, hindering real-world applications
- What if we had a **better translator**?



Introduction

Goal:

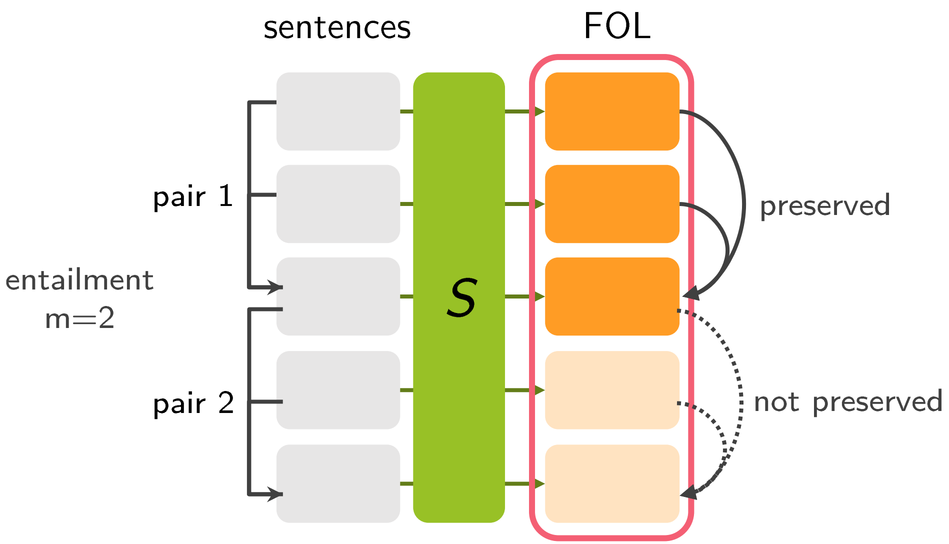
- We want a system that **translates NL to FOL**,
- so that the entailment in NL is **faithfully preserved in FOL space.**



Introduction

Metric: Entailment-Preserving Rate (EPR)

- Given a natural language entailment dataset $\mathcal{D} = \{((p_{i,1}, \dots p_{i,m_i}), h)\}_{i=1..N}$,
- Parse $p_{i,j}$ and h into FOL, **independently**
- Calculate the number of entailment-preserved instances among N.



→ **EPR = 1/2**

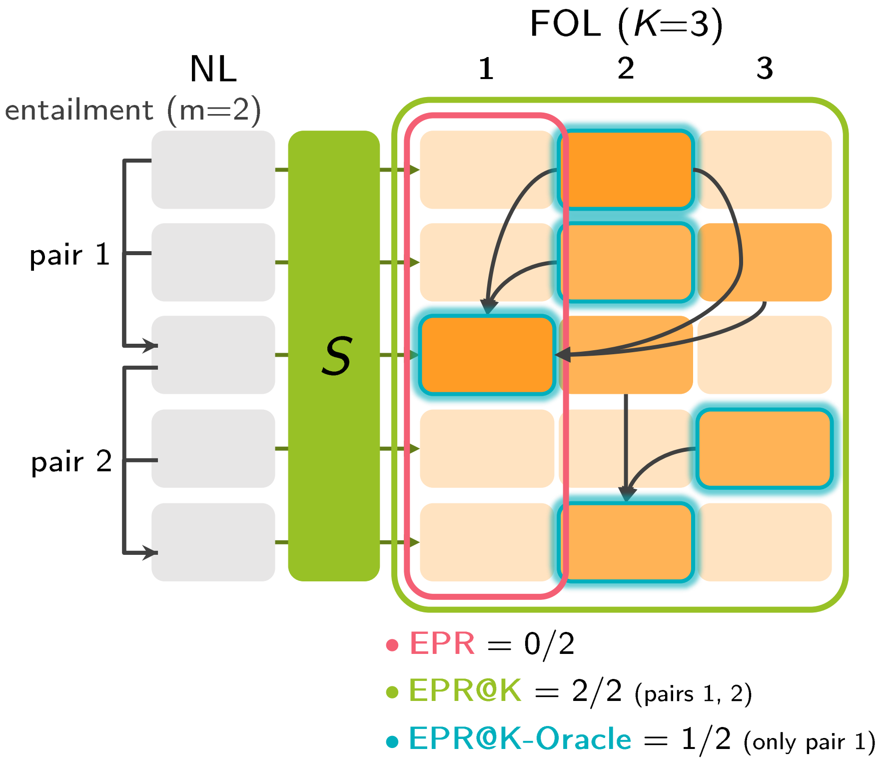
Introduction

Extensions of EPR

By sampling top K FOLs instead of 1, we can expand EPR to:

- EPR@K: If **any** of K^{m+1} combinations preserve entailment, it is a success
- EPR@K-Oracle: When selecting at most 1 FOL per each sentences, the max value of EPR

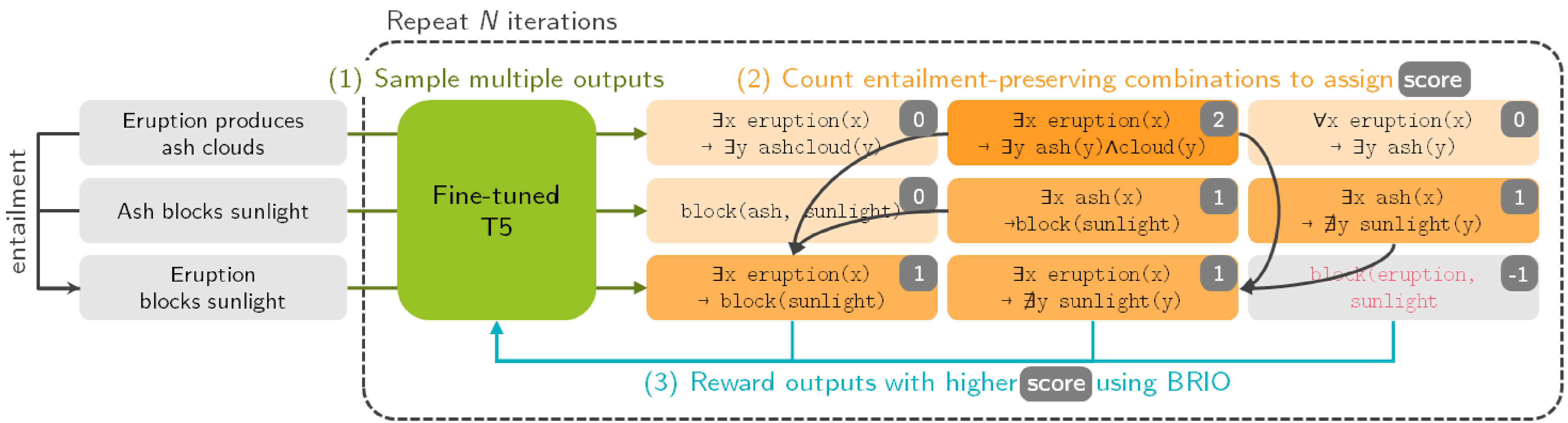
∴ EPR $<$ EPR@K-Oracle $<$ EPR@K



Method

Reinforcement Learning-like approach: Use the **natural entailment labels as the reward!**

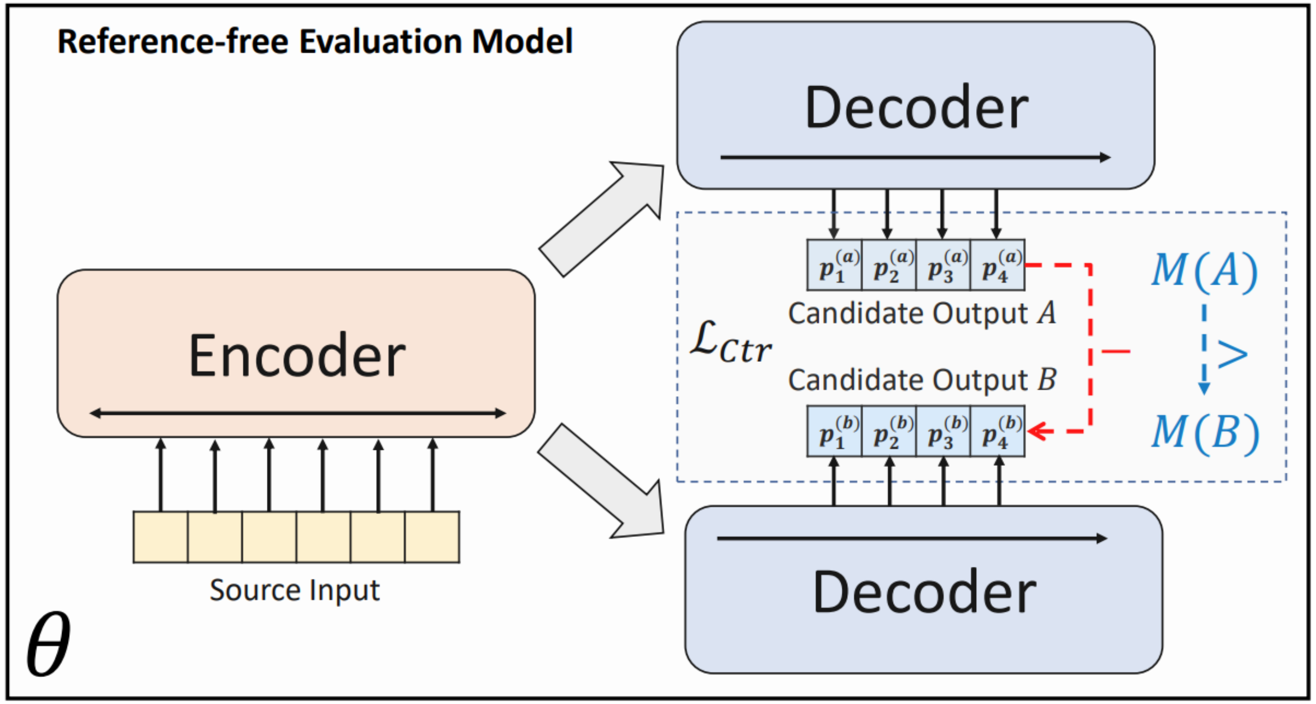
- Train *initial model* using parallel NL-FOL data
- Generate multiple FOL represetations for each NL sentence
- Reward combinations that preserve entailment, using RL-like objective (BRIO)
- Repeat the whole iteration multiple times



Method

BRIO: A **ranking loss** for seq2seq generation (Liu et al., 2022)

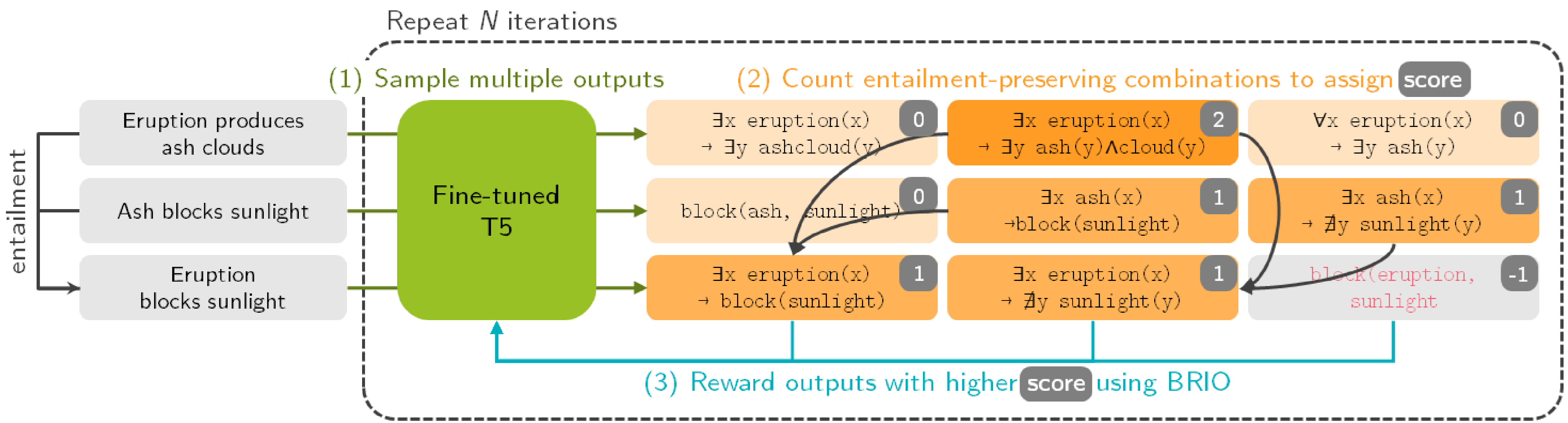
- Sample K outputs from a single input using the policy
- Rank K outputs based on external scoring function *s*
- Apply hinge (margin) loss to ensure that E(log$p(A)$) – E(log$p(B)$) $> \lambda$



Method

Reinforcement Learning-like approach: Use the **natural entailment labels as the reward!**

- Train *initial model* using parallel NL-FOL data
- Generate multiple FOL represetations for each NL sentence
- Reward combinations that preserve entailment, using RL-like objective (BRIO)
- Repeat the whole iteration multiple times



Experiments

Evaluation datasets: Three **multi-premise natural language entailment** datasets

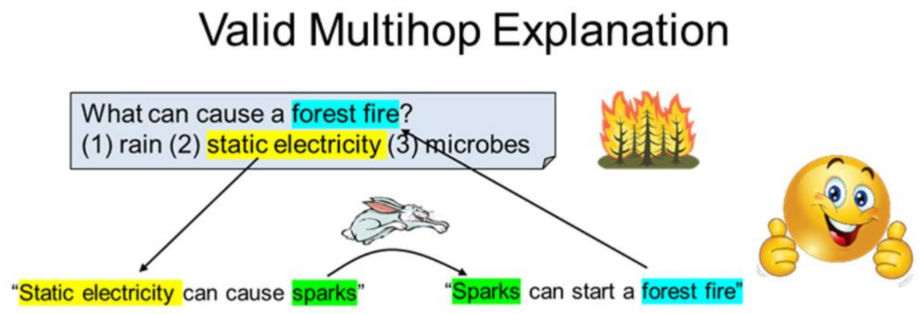
- e-SNLI, EntailmentBank, eQASC (clockwise from left-top)

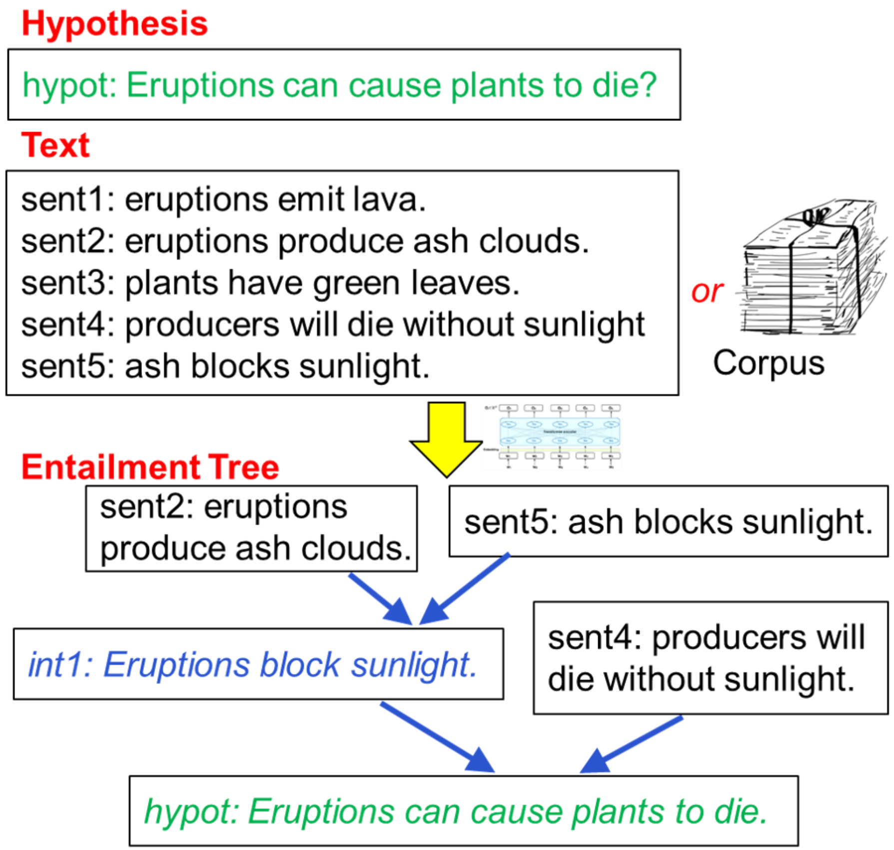
Premise: An adult dressed in black holds a stick.
Hypothesis: An adult is walking away, empty-handed.
Label: contradiction
Explanation: Holds a stick implies using hands so it is not empty-handed.

Premise: A child in a yellow plastic safety swing is laughing as a dark-haired woman in pink and coral pants stands behind her.
Hypothesis: A young mother is playing with her daughter in a swing.
Label: neutral
Explanation: Child does not imply daughter and woman does not imply mother.

Premise: A man in an orange vest leans over a pickup truck.
Hypothesis: A man is touching a truck.
Label: entailment
Explanation: Man leans over a pickup truck implies that he is touching it.

Figure 1: Examples from e-SNLI. Annotators were given the premise, hypothesis, and label. They highlighted the words that they considered essential for the label and provided the explanations.





Experiments

Fine-tuning corpora: MALLS (Yang et al., 2024)

- NL<>FOL parallel corpus generated by GPT-4
- Used to fine-tune our initial model & baselines

MALLS	**NL:** A car must have a motor and wheels to be considered functional. **FOL:** $\forall x\ \texttt{Car}(x) \land \texttt{Functional}(x) \rightarrow (\texttt{HasMotor}(x) \land \texttt{HasWheels}(x)))$. **Error:** none
	NL: A grocery store sells food and household items. **FOL:** $\forall x \exists y \exists z (\texttt{GroceryStore}(x) \land \texttt{Food}(y) \land \texttt{HouseholdItem}(z) \land \texttt{Sells}(x, y) \land (\texttt{Sells}(x, z))$. **Error:** none

Baselines: Existing methods for translating NL to FOL

- Semantic parse-based translators (2010-2018)
  - NL sentences $\rightarrow$ Neural AMR parser $\rightarrow$ Rule-based translation to FOL
- End-to-end neural models (2022-2024)
  - LLaMA / T5 fine-tuned on MALLS
  - GPT-4o / GPT-4o-mini (5-shot)

Results

How much portion of natural entailment can the translator preserve?

Outperforms both syntax-based methods and end-to-end generative models

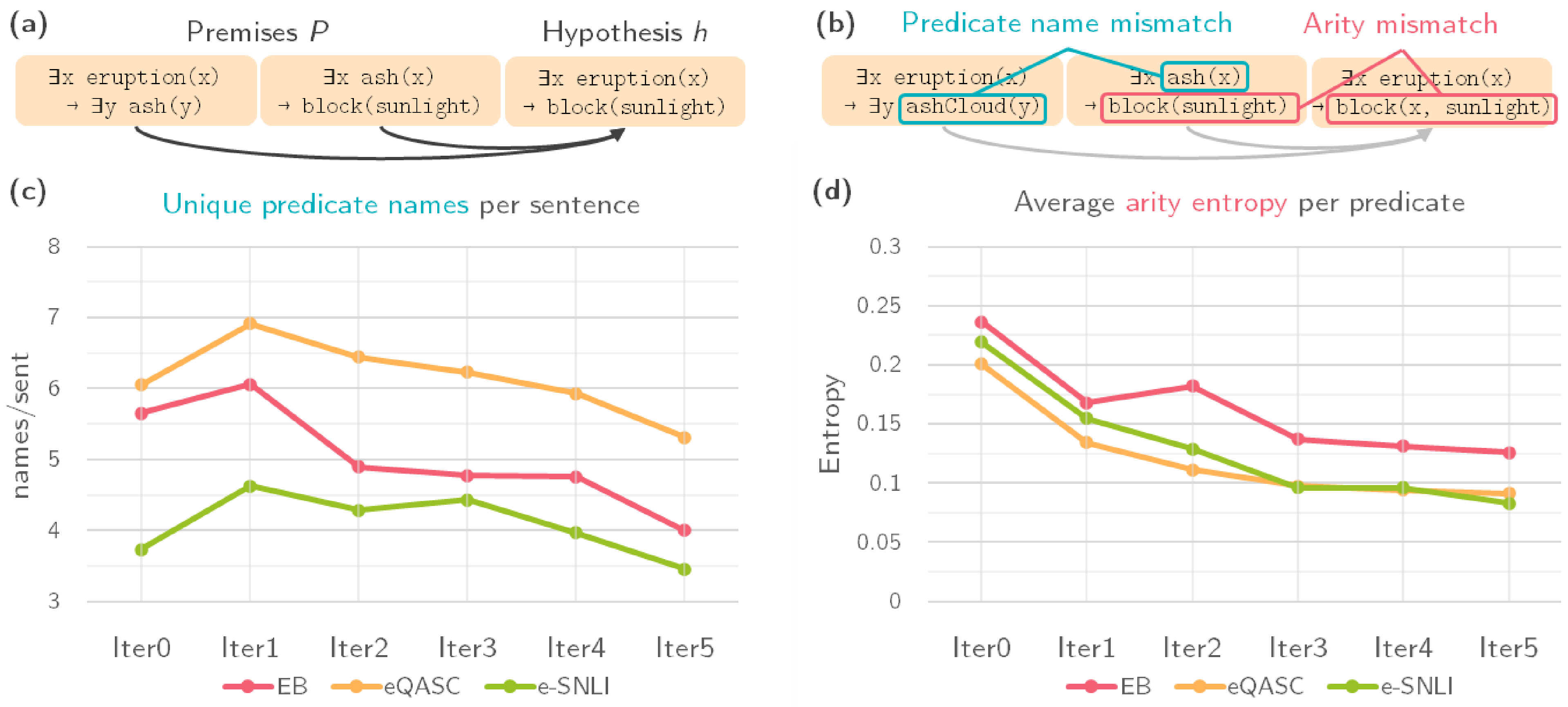
Metric	Method	EB	eQASC	e-SNLI
EPR	CCG2Lambda	0.0	0.0	0.0
	AMR2FOL(Bos)	0.0	0.0	2.5
	AMR2FOL(Lai)	0.0	0.0	1.6
	GPT-4o-mini	3.2	2.4	0.9
	GPT-4o	2.9	1.1	1.5
	LogicLLaMA	5.2	2.5	0.7
	T5-Iter0	5.6	2.6	0.1
	T5-Iter5	**7.4**	**4.9**	**4.3**
EPR@16	GPT-4o-mini	10.5	7.6	8.3
	GPT-4o	13.2	11.4	8.3
	LogicLLaMA	5.2	2.5	0.7
	T5-Iter0	15.4	12.5	3.4
	T5-Iter5	**32.8**	**33.1**	**36.1**
EPR@16 Oracle	GPT-4o-mini	10.5	7.4	5.6
	GPT-4o	13.0	10.8	5.6
	LogicLLaMA	5.2	2.5	0.7
	T5-Iter0	15.2	11.7	0.1
	T5-Iter5	**31.1**	**28.3**	**24.0**

Table 2: EPR, EPR@16, and EPR@16-Oracle measured on three different datasets (EntailmentBank (EB), eQASC, e-SNLI), single-run.

Results

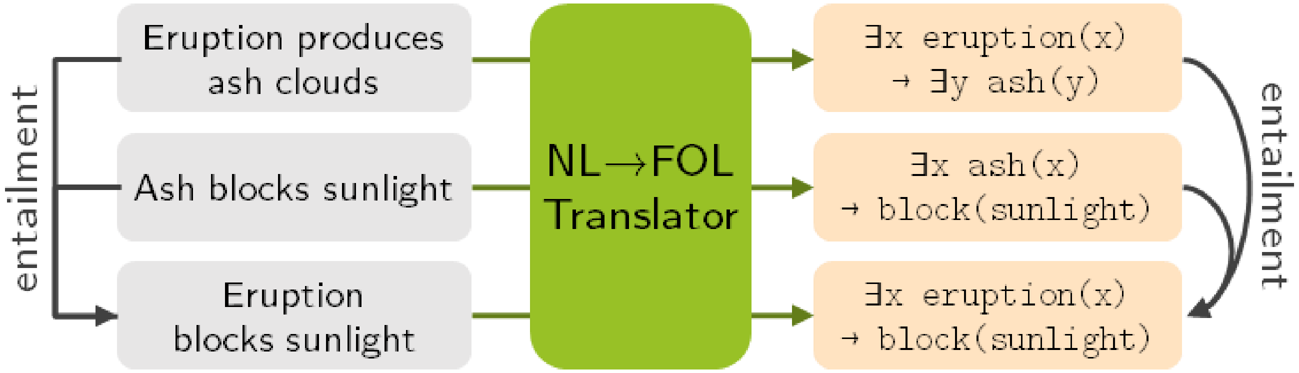
Predicate name/arity mismatch leads to failure in preserving entailment

$\rightarrow$ Model learns to use **unified predicate signatures** across sentences, reducing arbitrarines



Conclusion

Trained a translator from NL to FOL based on **distant entailment labels**



Semantic parser **can be trained** by reasoning execution results

- Highly similar to repo-level code generation / multi-intent NL2SQL parsing / …

Summary

- Logic is a power tool for solving natural language reasoning problems
- Interaction between semantic parsing (NL$\rightarrow$Logic) and execution (prover) is important
- Modeling the interaction is crucial for developing versatile neuro-symbolic reasoner
  - **Interleaving** semantic parsing and execution
    - *Work 1: Symbolic Backward Chaining*
  - Using desired execution results as **training objective** for parsers
    - *Work 2: Entailment-preserving FOL representations*

Future works

- Neuro-symbolic reasoning in more complex scenario (Olympiad-level math, law, medical, …)
- Can neuro-symbolic reasoning be used as a *teacher/reward model* for strong LLMs?