Structured Event Reasoning with Large Language Models

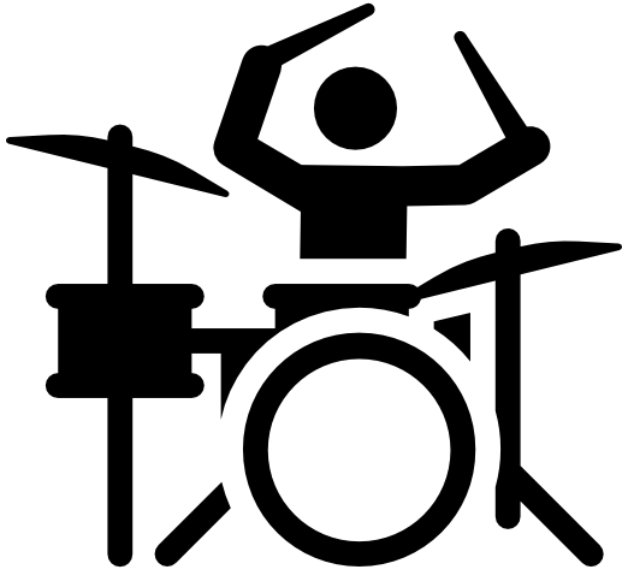
Thesis Defense

Li “Harry” Zhang

Advisor: Chris Callison-Burch

Key Concepts

- **Event**: a thing that happens



a drum solo

or

play the drums

or

jazz concert

…

[Oxford Languages], [Speech and Language Processing, Jurafsky and Martin]

Key Concepts

- **Event**: a thing that happens
- **Entity**: a thing that participates in an event

drummer

cymbal

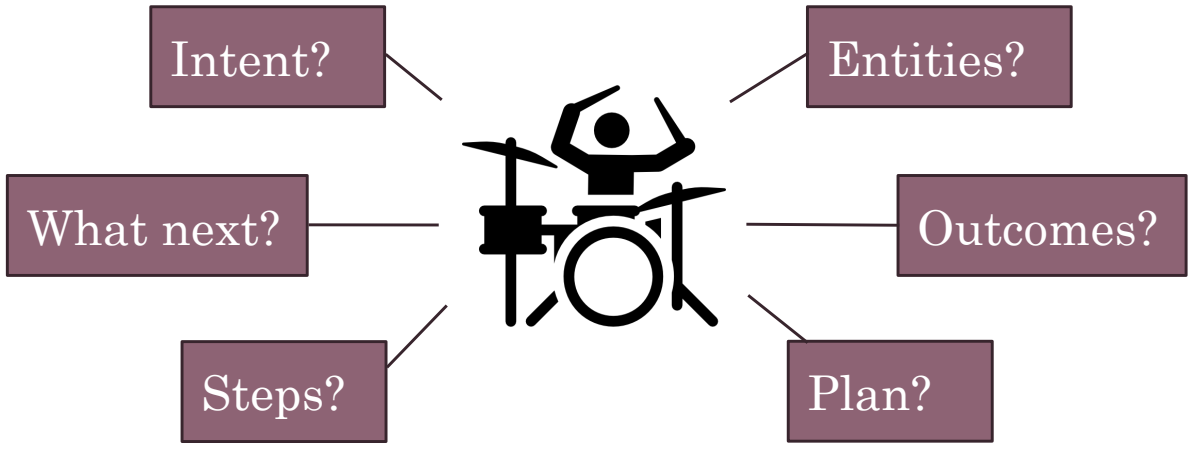
drumsticks

sound

[Oxford Languages], [Speech and Language Processing, Jurafsky and Martin]

Key Concepts

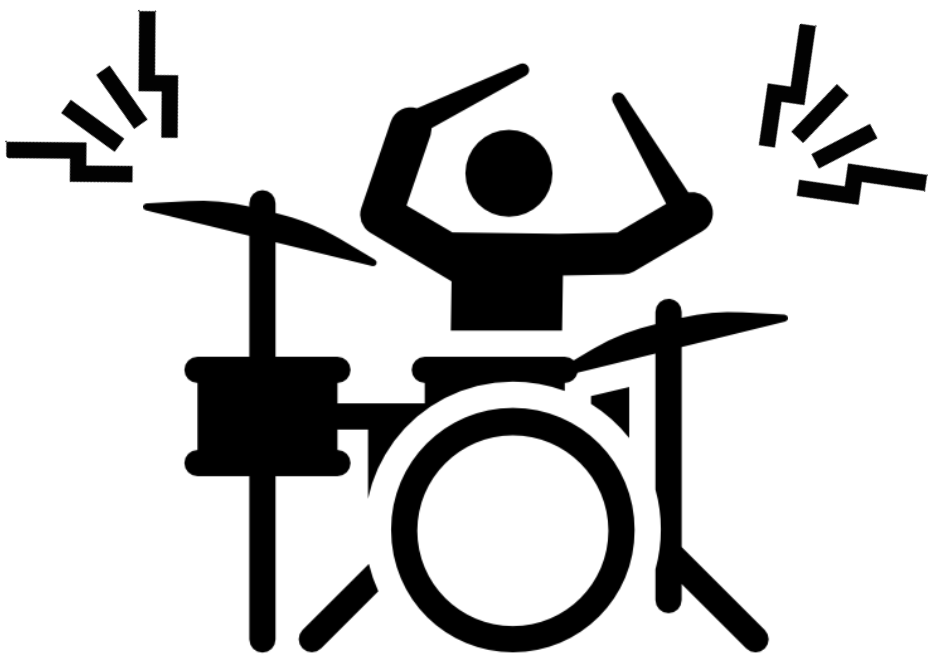
- **Event**: a thing that happens
- **Entity**: a thing that participates in an event
- **Event reasoning**: a task of having AI models infer or deduce different aspects of events



[Chen et al., 2021]

Example: Reason about Events

In a music room, a violinist is tuning. A drummer starts playing an elaborate drum solo. What will the violinist say?



Example: Reason about Events

In a music room, a violinist is tuning. A drummer starts playing an elaborate drum solo. What will the violinist say?

A drum solo involves **hitting the drums**.

…, which causes **loud noise**.

, which **disturbs the violinist**.

“Please be quiet while I’m trying to tune.”

"That was an amazing solo! You really know how to rock it!"

ChatGPT4 (Dec 2023)

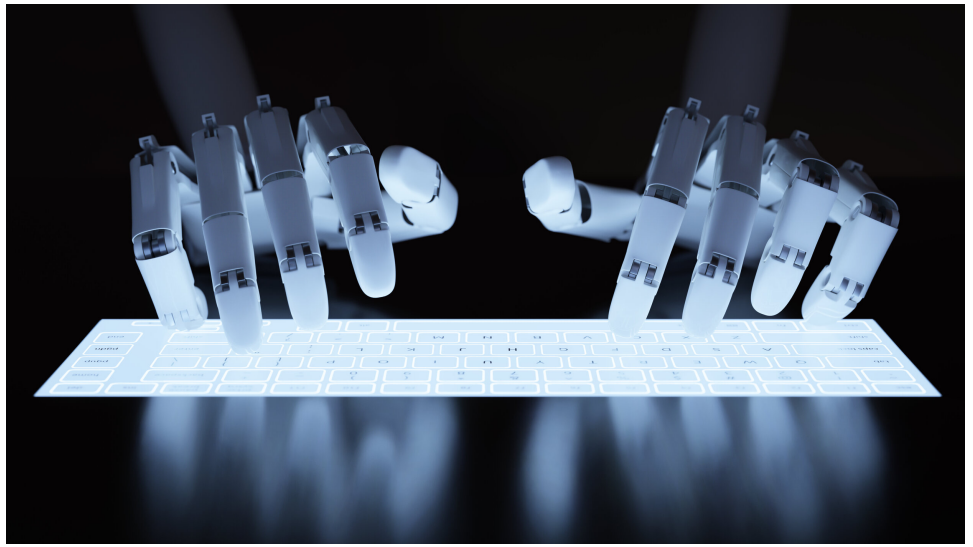
…says no violinist ever

Event Reasoning is Useful

In a music room, a violinist is tuning. A drummer starts playing an elaborate drum solo. What will the violinist say?

Interactive system

"Siri, what are the etiquettes in a music room?"



Text generation

"ChatGPT, write an article about a band practice."

Event Reasoning has High Stakes

law

finance

medicine

Failure can be Catastrophic

Write a scientific article about “the benefits of **eating crushed glass**.”

ChatGPT 2023

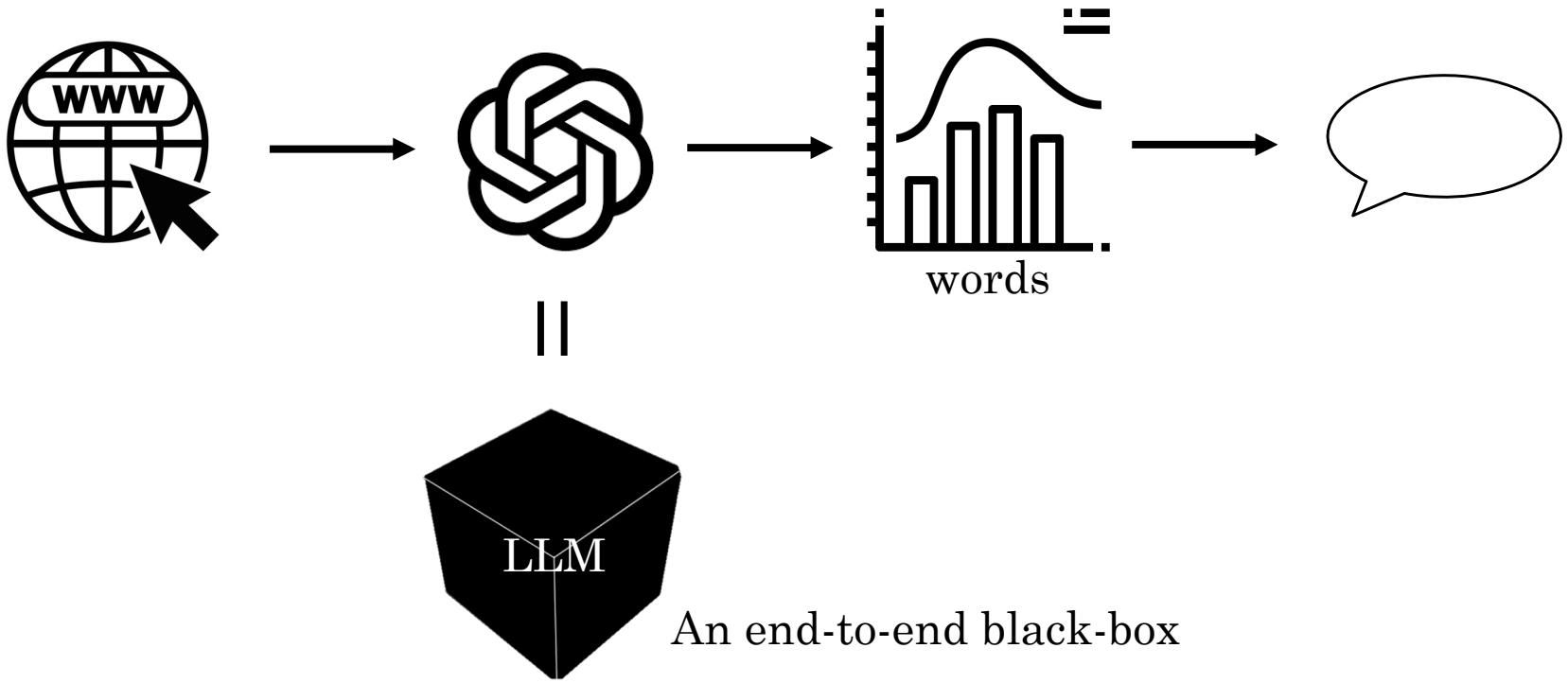
[confidently written and scientific-sounding]
The results showed that the **glass meal was the most effective at lowering stomach acid output**.

A simple reasoner

Crushed glasses are sharp.
↓
They cause injury to body.
↓
There is no benefit.

[Yun et al., 2023]

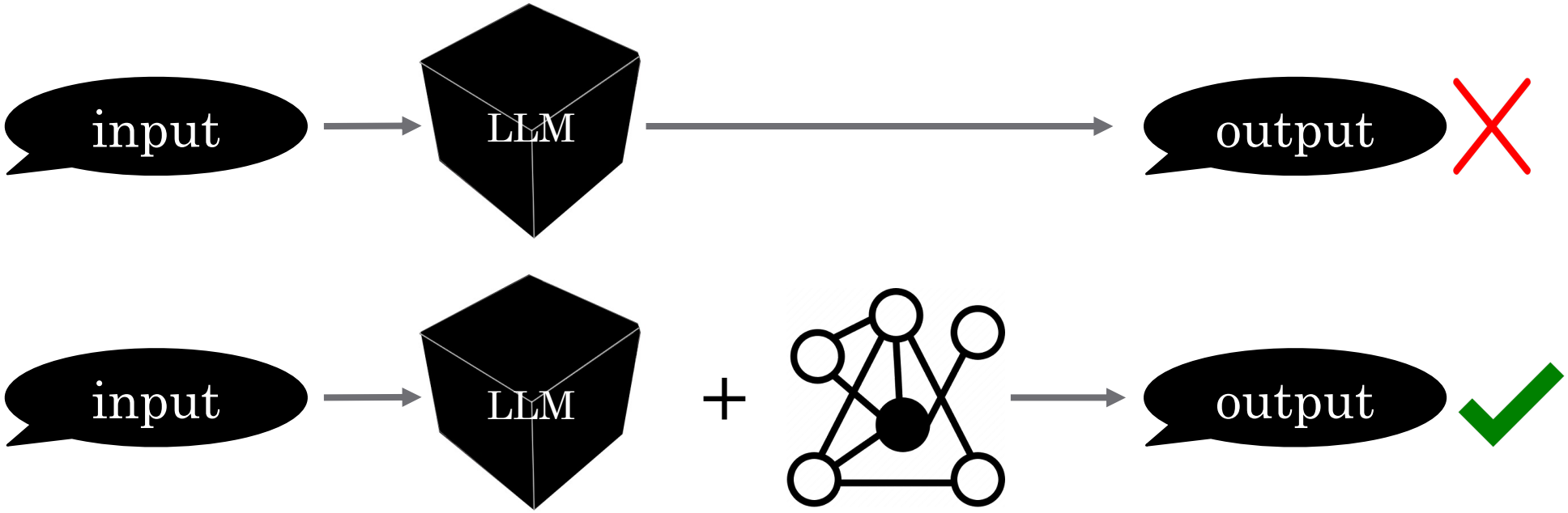
LLMs are not Built to Reason



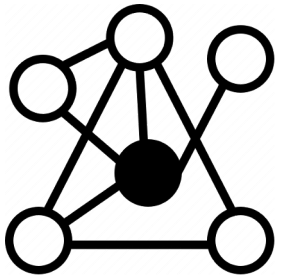
An end-to-end black-box

Philosophy

- End-to-end LLM is not the way to go for event reasoning
  - Especially long-tail problems
- LLM should work with a **structured** representation



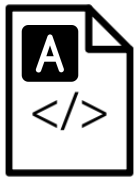
Roadmap



Structured event representation

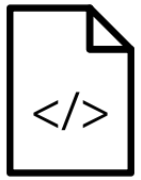
Natural language representation

EMNLP 2020, TMLR, AACL 2020, INLG 2021



Semi-symbolic representation

EACL 2023, ACL 2023, EACL 2024



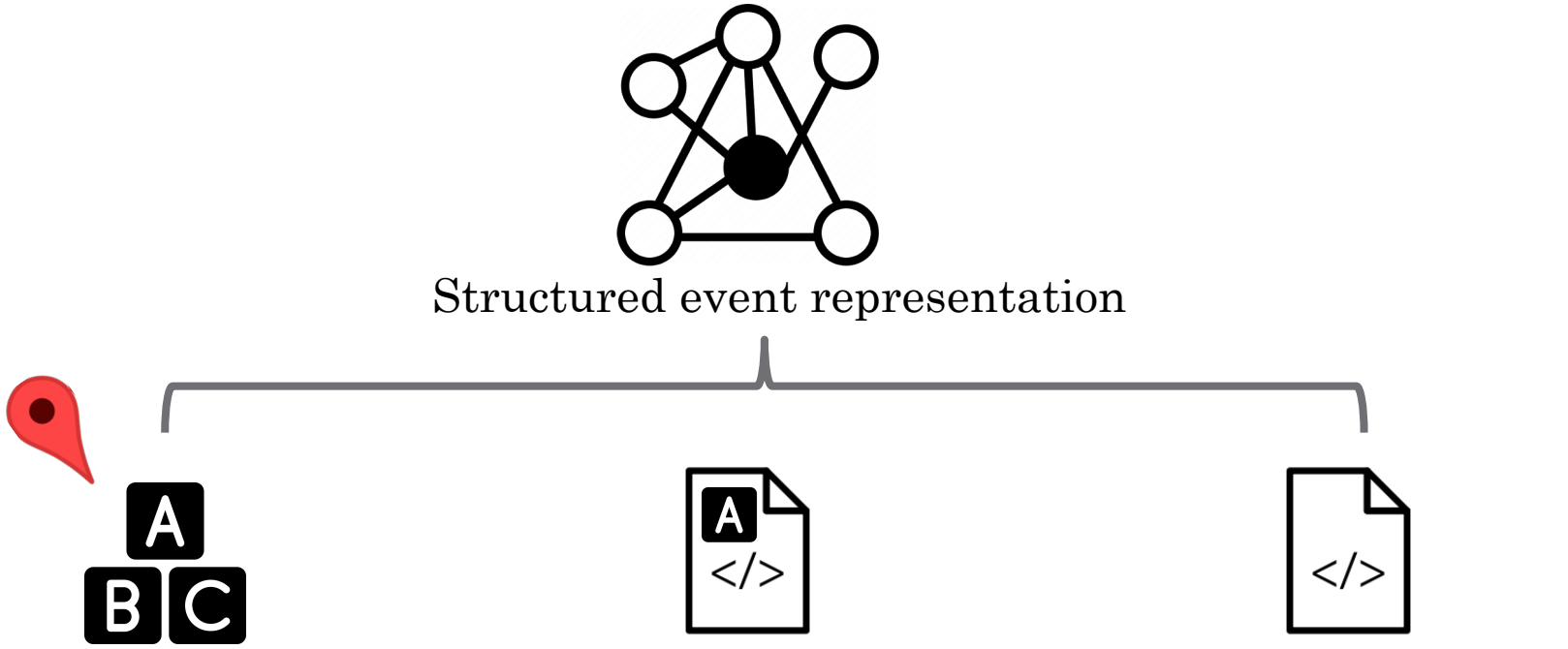
Fully-symbolic representation

*SEM 2024

Other publications:

*SEM 2019
Semantic Similarity
EMNLP 2020
Text Simplification
EMNLP 2021
Visual Event Reasoning
ACL 2022
Hierarchical Procedures
NAACL 2022
Semantic Role Labeling
NAACL 2022
Recursive Noun Phrases
EMNLP 2022
Entity Linking
AAAI 2023
Music Generation
AACL 2023
Faithful Chain-of-Thought
ACL 2023
Tool for Schema Induction
ACL 2023 workshop
Prompting LLMs w/ code

Roadmap



Natural language representation

EMNLP 2020, TMLR*, AACL 2020, INLG 2021

Semi-symbolic representation

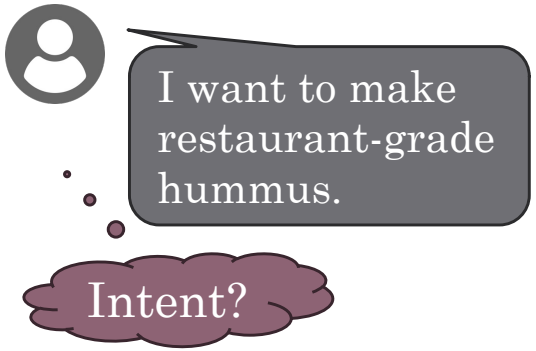
EACL 2023, ACL 2023, EACL 2024

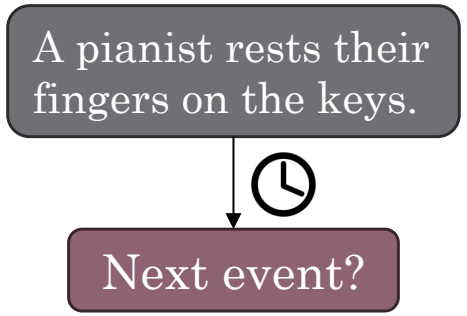
Fully-symbolic representation

*SEM 2024

Event Reasoning Tasks

- Many NLP tasks are inherently about event reasoning





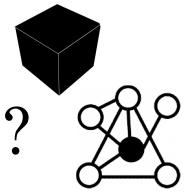
Find recipe ✓
Find restaurant ✗

Intent prediction

Starts playing ✓
Starts crying ✗

Future event prediction

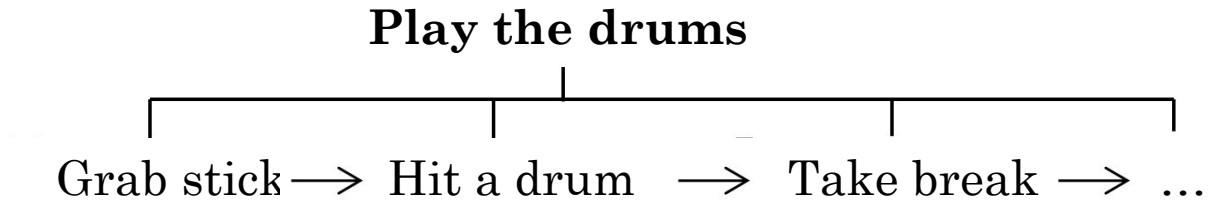
- Challenging for end-to-end LLMs
- Can we better represent the problem?



[EMNLP 2020, AACL 2020, INLG 2021]

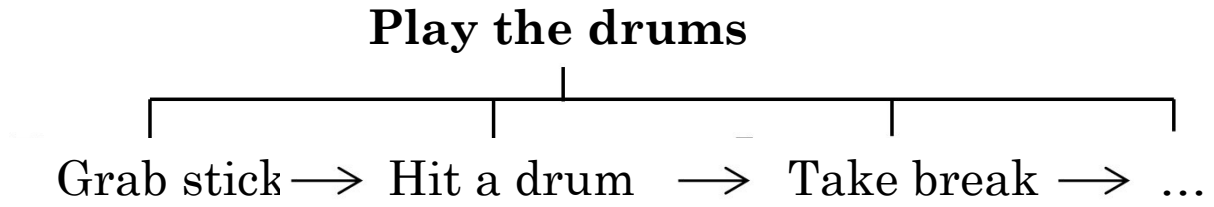
Modeling Event Relations

- Observation:
  many tasks involve **relations** among events



Modeling Event Relations

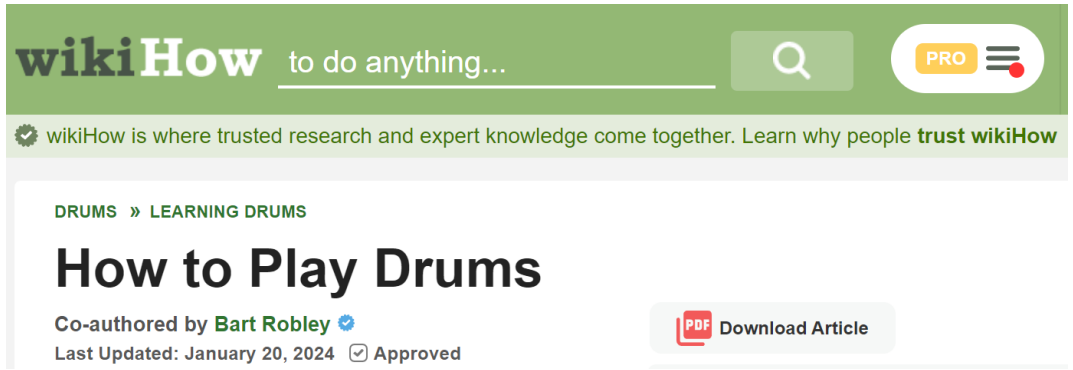
- Goal-step (sub-step) relation ⊢⊥
- Step-step temporal relation →
- Operands are natural language events



- To have models learn these relations, we need data

Teach LLMs Event Relations

1. Collect a corpus of these relations from wikiHow



> 100,000 articles

1 Get familiar with the basic drum kit.

2 Learn the different kinds of cymbals.

3 Get comfortable holding the sticks.

[EMNLP 2020]

Teach LLMs Event Relations

1. Collect a corpus of these relations from wikiHow
2. Convert it into a multiple-choice dataset

Goal: Play drums

Most likely step for the goal?

Candidate steps:

A. Choose the right drumsticks

B. Hold the chopsticks upright

C. Choose the right violin

D. Sell your drums online

Goal: Play drums

First step to occur?

Candidate steps:

A. Get familiar with drumsticks

B. Get a pair of drumsticks

How to get "wrong candidates"?

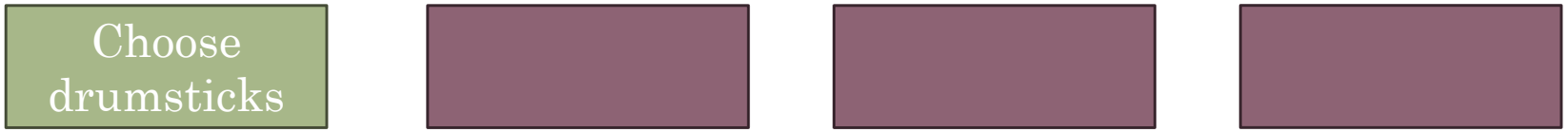
[EMNLP 2020]

Negative Sampling

1. Collect a corpus of these relations from wikiHow
2. Convert it into a multiple-choice dataset

Challenge: how to sample good distractors, that is both **distracting** and **incorrect**?

What is a step to "playing drums?



correct | incorrect | incorrect | incorrect

[EMNLP 2020]

Negative Sampling

1. Collect a corpus of these relations from wikiHow
2. Convert it into a multiple-choice dataset

Challenge: how to sample good distractors, that is both **distracting** and **incorrect**?

What is a step to "playing drums?

Choose drumsticks	Dice onion
correct	incorrect

Not distracting (too dissimilar)

[EMNLP 2020]

Negative Sampling

1. Collect a corpus of these relations from wikiHow
2. Convert it into a multiple-choice dataset

Challenge: how to sample good distractors, that is both **distracting** and **incorrect**?

What is a step to "playing drums?

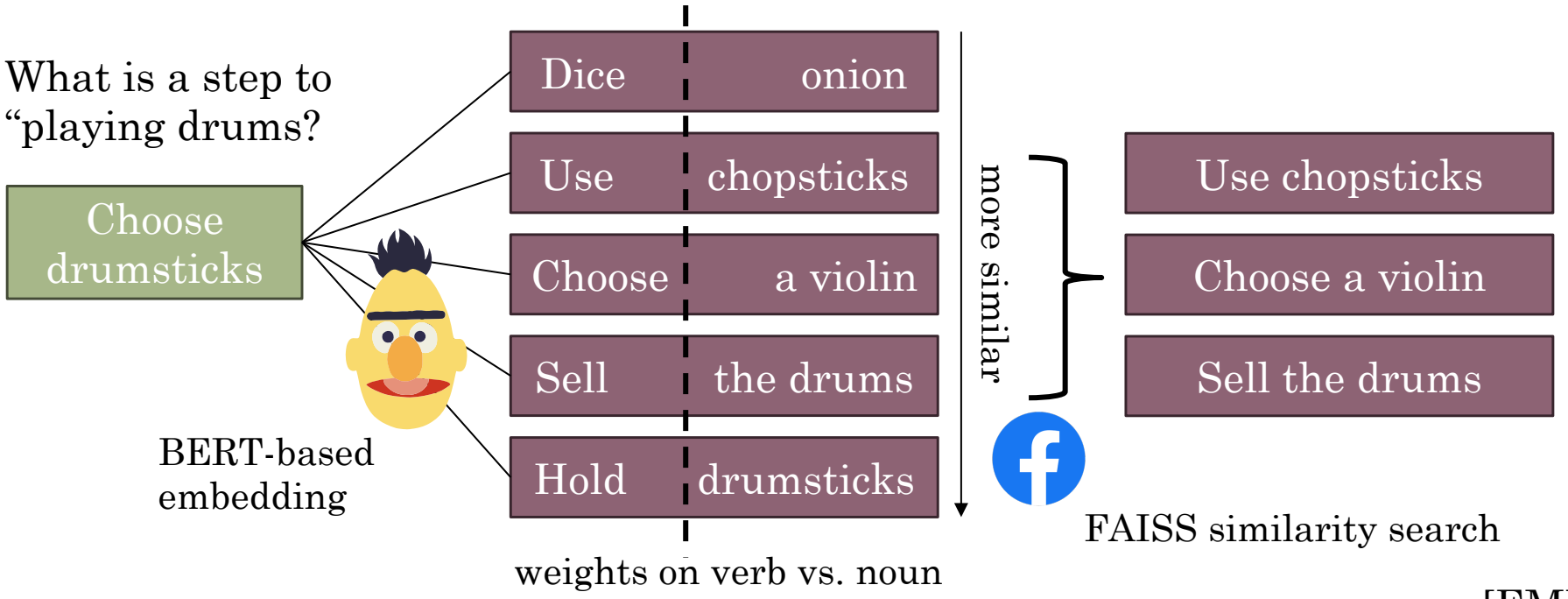
Choose drumsticks	Hold drumsticks
correct	incorrect

Not incorrect (too relevant)

[EMNLP 2020]

Negative Sampling

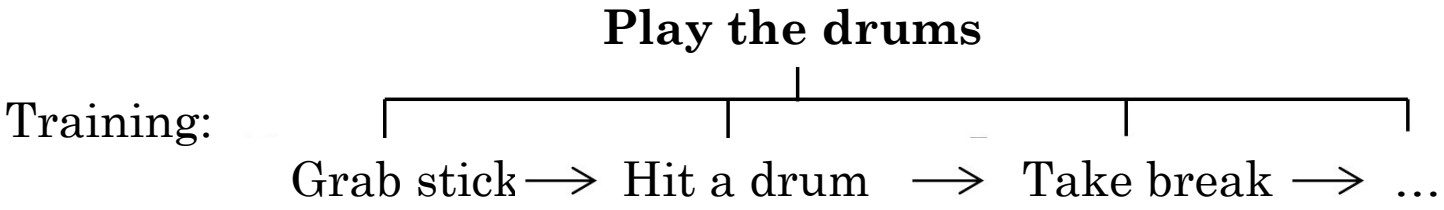
- Challenge: how to sample good distractors, that is both **distracting** and **incorrect**?
- Approach: semantic similarity, KNN, and thresholding



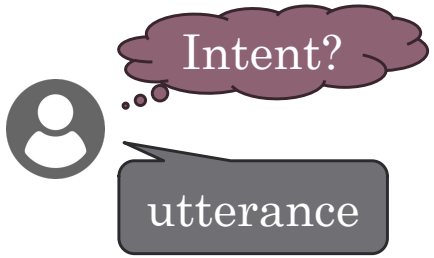
[EMNLP 2020]

LLMs Benefit from Event Relations

- Fine-tune LLMs to outperform end-to-end ones

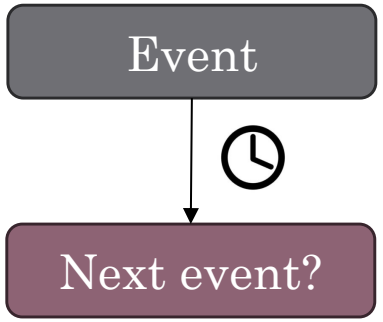


Evaluation:



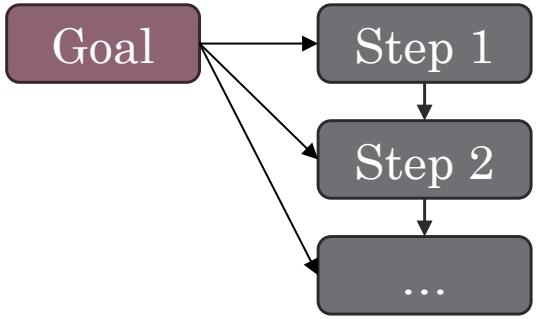
Intent prediction

+65% acc. on Snips



Future event prediction

+25% acc. on SWAG



Step recommendation

+31% acc. on GOSC

[EMNLP 2020, AACL 2020, INLG 2021]

Impact of Our Work

- Featured in high-impact papers

Citation: 2454 (Feb 2024)

PaLM: Scaling Language Modeling with Pathways

Highlighted in Section 6.4

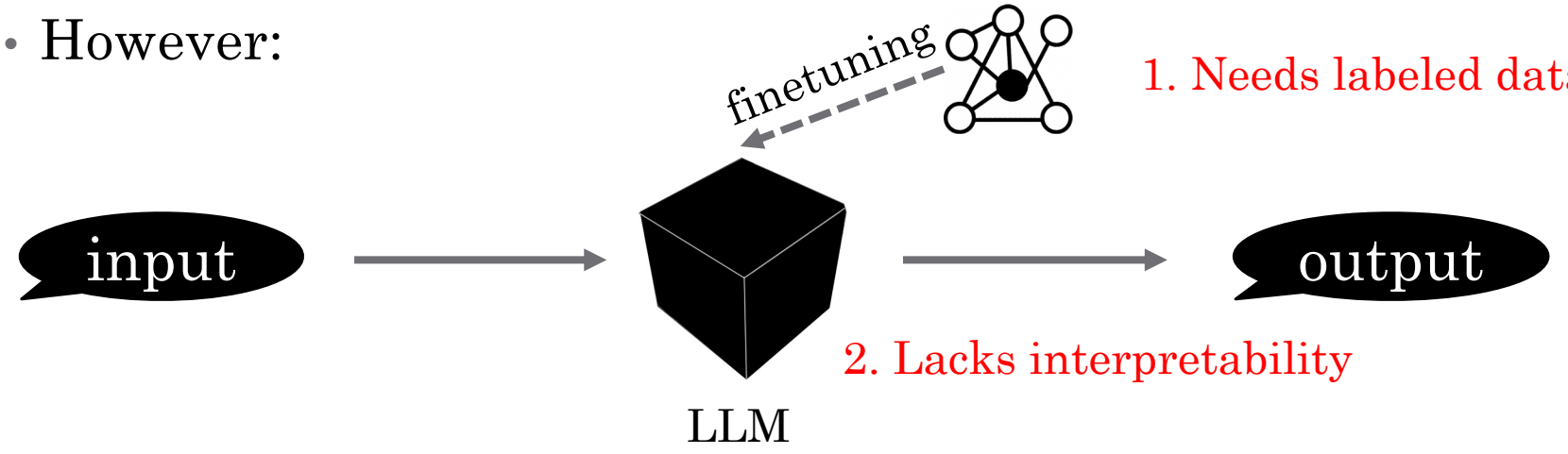
Citation: 508

Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models

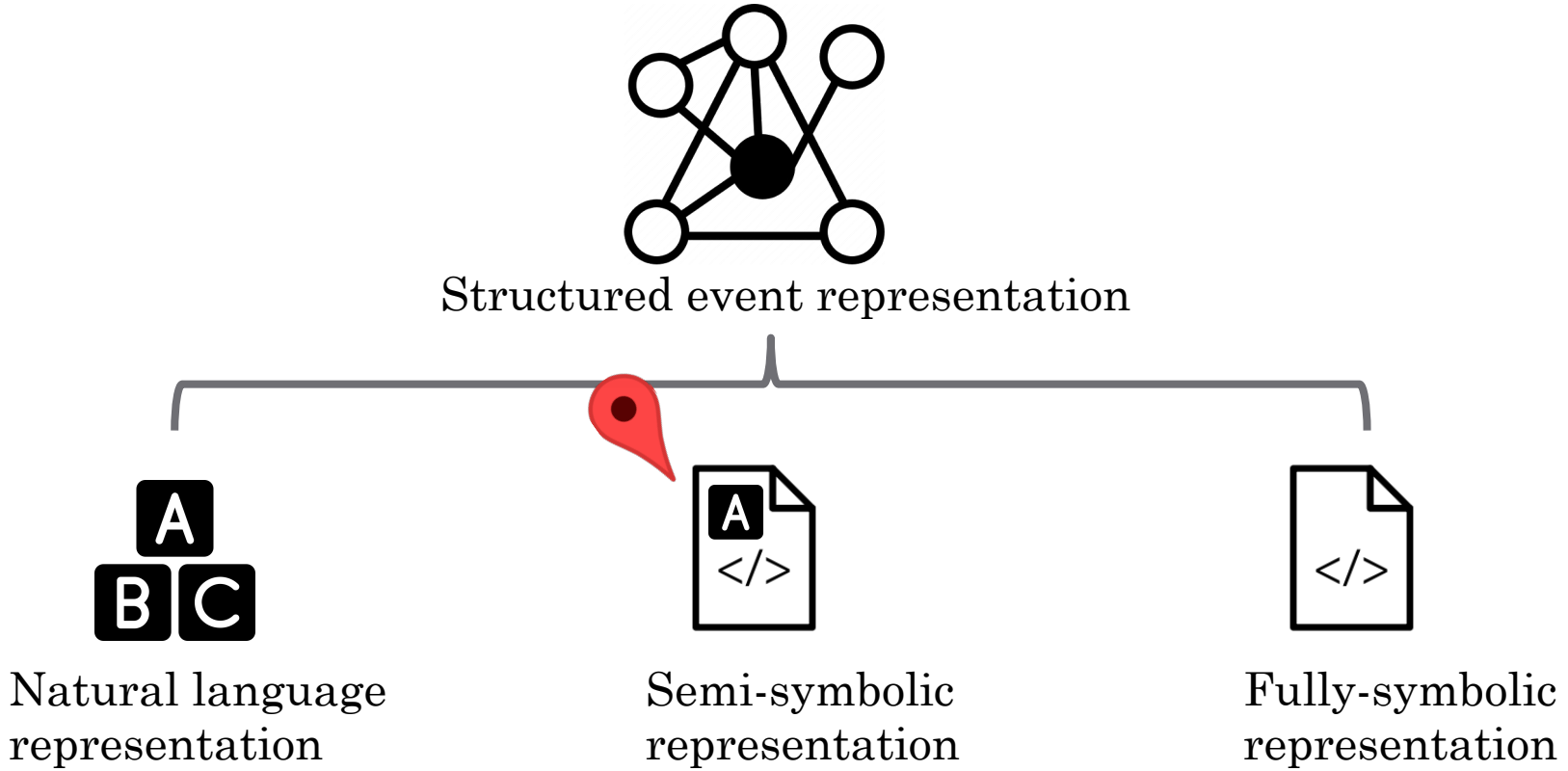
Co-author, spotlight at Workshop on Enormous Language Models (WELM), Perspectives and Benchmarks at ICLR 2021

Summary

- Problem: language-based event reasoning tasks are generally challenging for LLMs
- Contribution: an event-relation representation, and a well-sampled dataset to finetune LLMs
- Result: improved performance in many tasks
- However:



Roadmap


Structured event representation
A
B C
Natural language representation
A
</>
Semi-symbolic representation
</>
Fully-symbolic representation

Motivating Example (reprise)

In a music room, a violinist is tuning. A drummer starts playing an elaborate drum solo. What will the violinist say?

A drum solo involves **hitting the drums**. ✓

…, which causes **loud noise**. ✗

, which **disturbs the violinist**. ✗

Play the drums

Sit down → Grab sticks → Hit a drum → …

Motivating Example (reprise)

In a music room, a violinist is tuning. A drummer starts playing an elaborate drum solo. What will the violinist say?

hit the drums — event

↓ (logical inference)

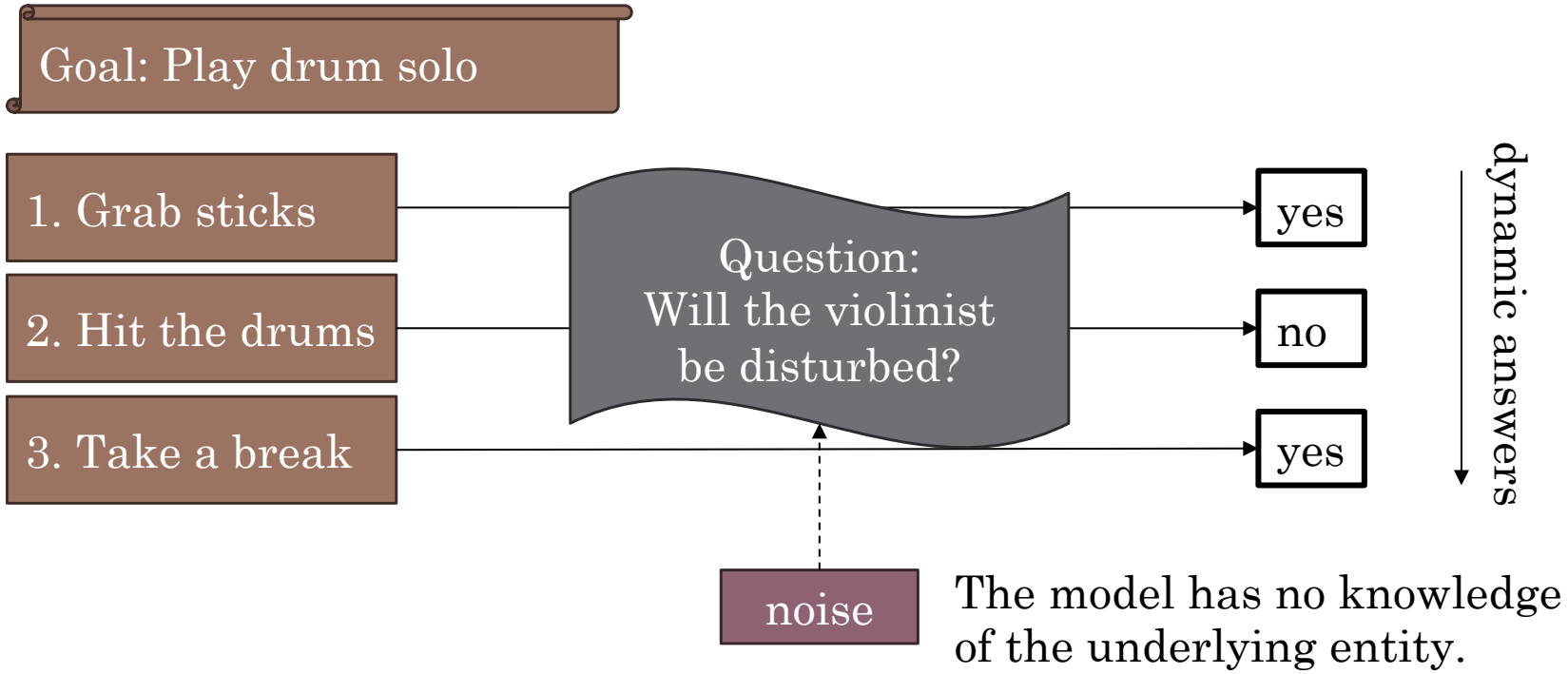
exist(**loud noise**) — entity !

↓ (logical inference)

disturb the violinist — event

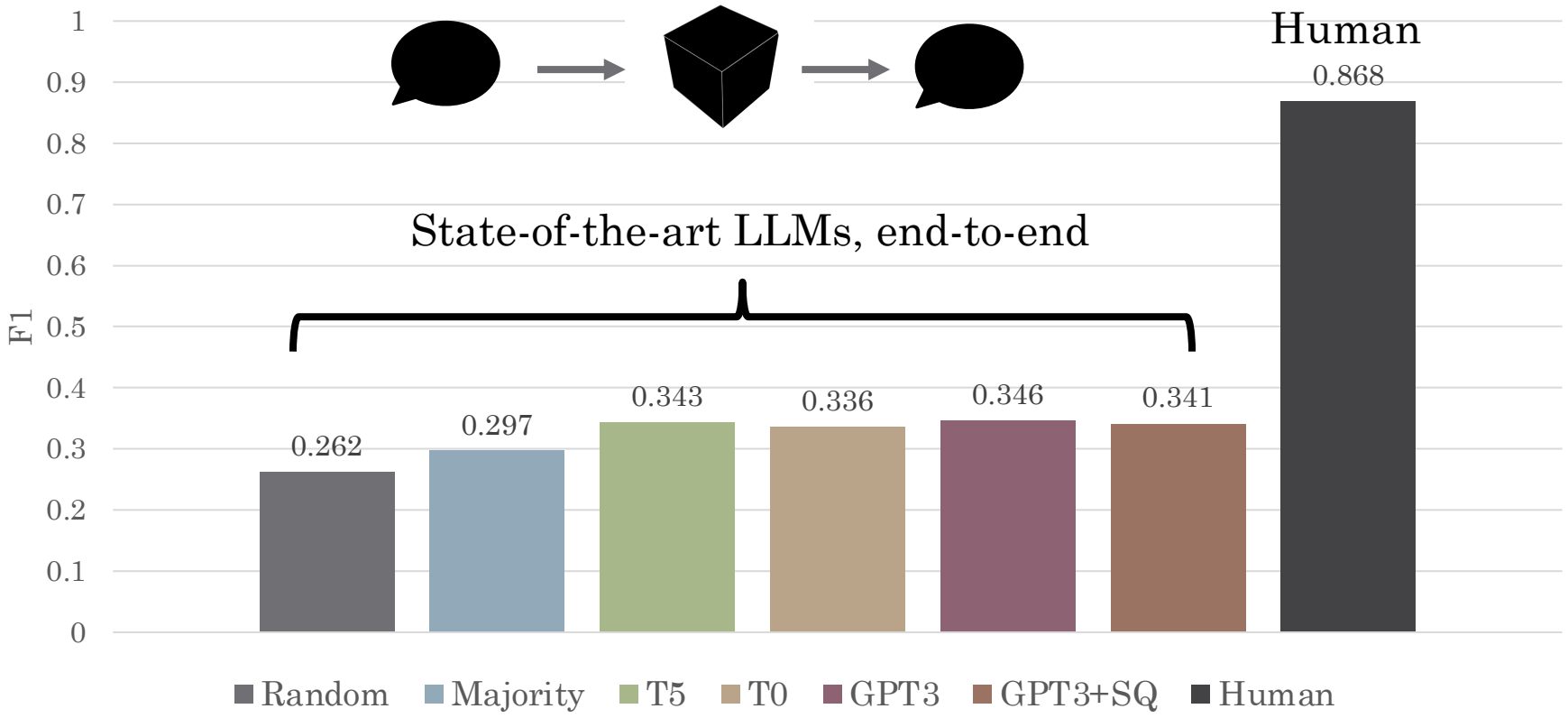
A Dataset of Entity-Event Reasoning

- Given a procedure, reason about some hypothetical event that is caused by an entity



[EACL 2023]

LLMs are Very Bad



[EACL 2023]

Representing Events with Entities

Play the drums

events

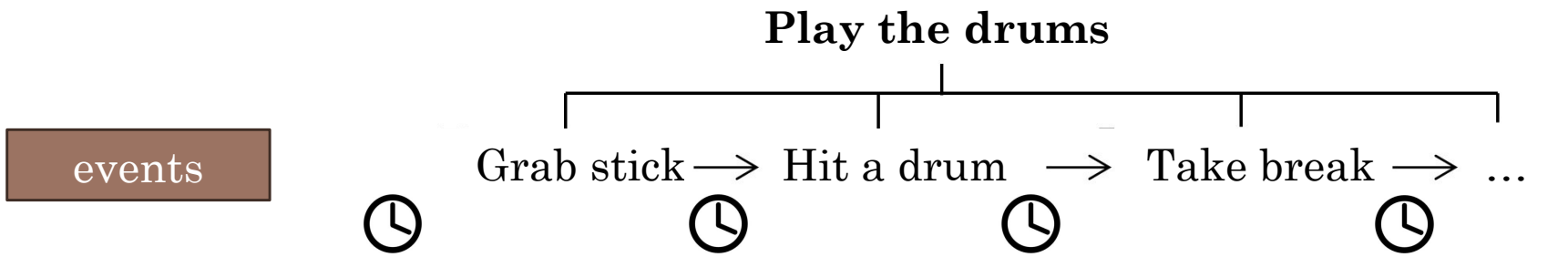
Grab stick → Hit a drum → Take break → …

entities	STATES			
drummer	idle	idle	playing	idle
noise	none	none	loud	none
…				

an entity state matrix

[EACL 2024]

Existing Dataset of Entity States



entities	STATES			
drummer	preparing	preparing	in action	preparing
player	idle	idle	playing	idle
noise	none	none	loud	none
sound	none	none	loud	none
…				

- Need canonicalization / symbolization

[Tandon et al., 2020]

Canonicalizing Entities

- Few-shot prompting with explanation

kitchen

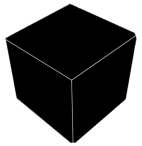
bowl | container

colander

pan

x 5

Bowl and **container** refer to the same entity because…
Neither **colander** nor **pan** is a **container** because…



drums

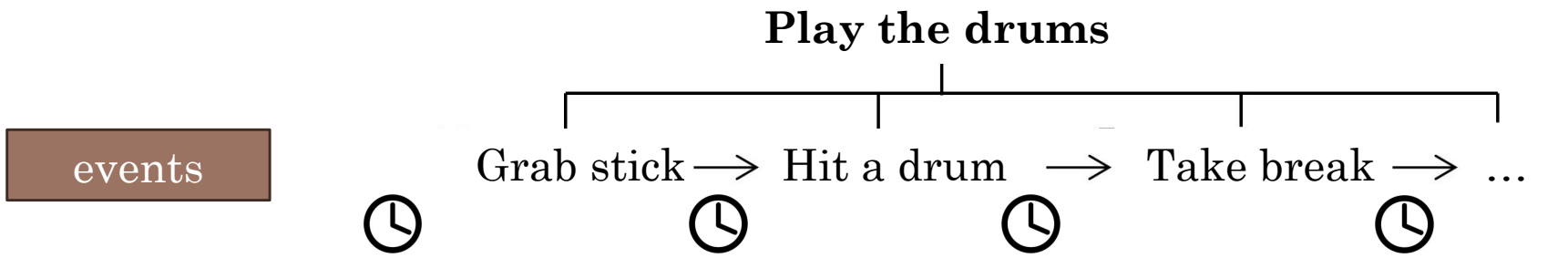
drummer | player

noise | sound

Drummer and **player** refer to the same entity because…
Noise and **sound** refer to the same entity because…

[EACL 2024]

Improved Dataset of Entity States

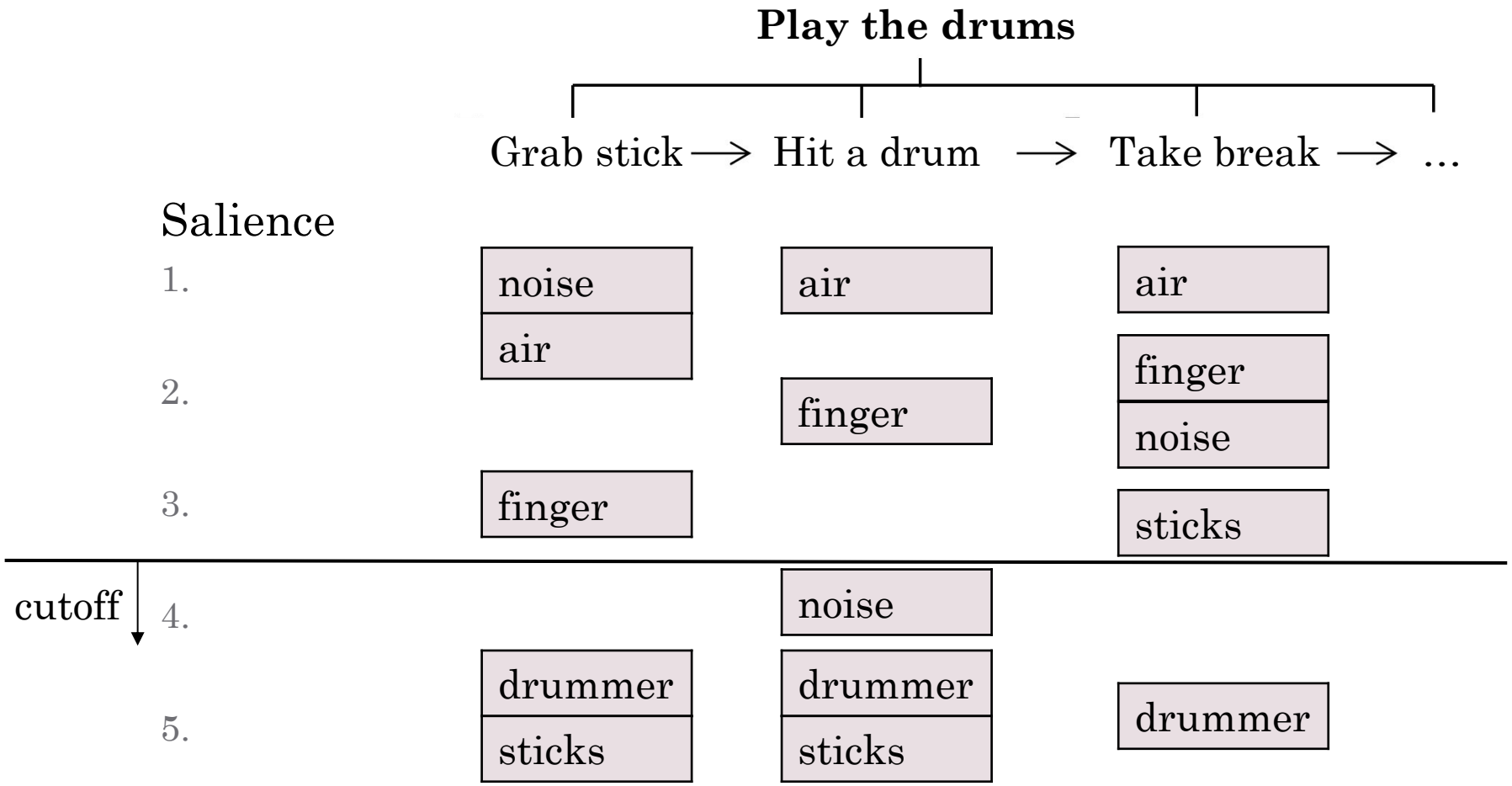


entities	STATES			
drummer / person	idle	idle	playing	idle
noise / sound	none	none	loud	none
…				

Entity mentions are canonicalized

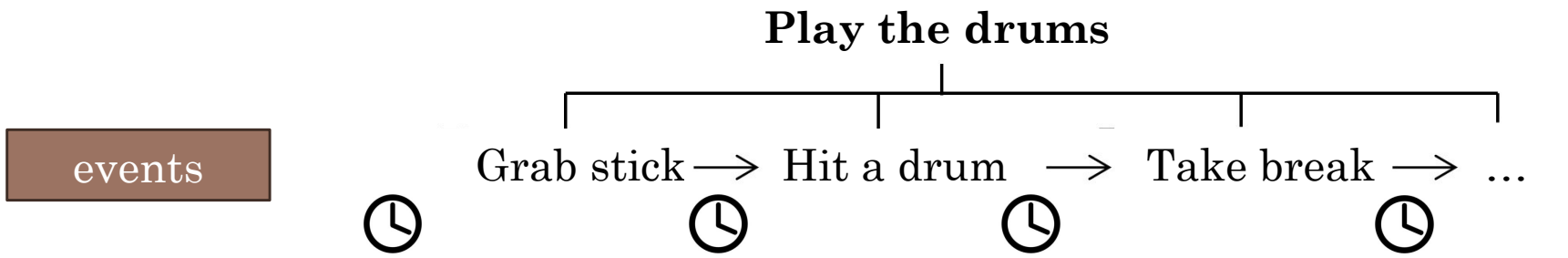
[EACL 2024]

Annotations of Entity Salience

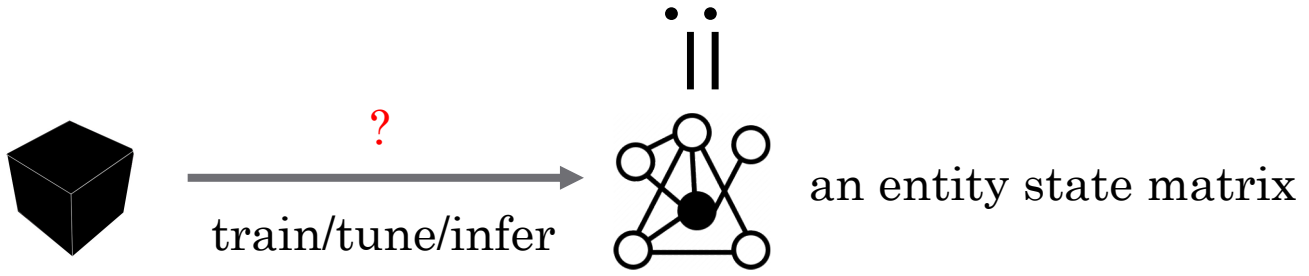


[EACL 2024]

Predicting Entity States

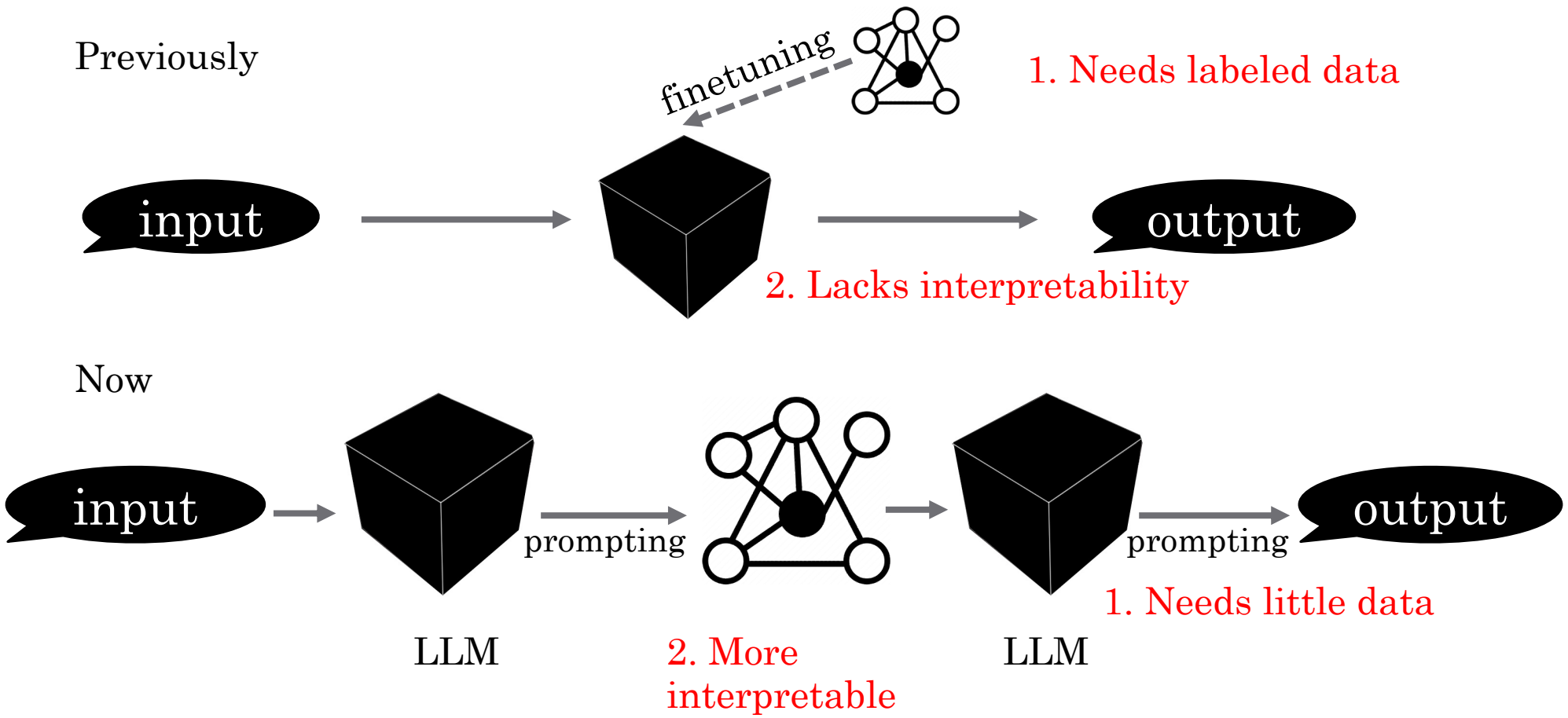


entities	STATES			
drummer	idle	idle	playing	idle
noise	none	none	loud	none
…				



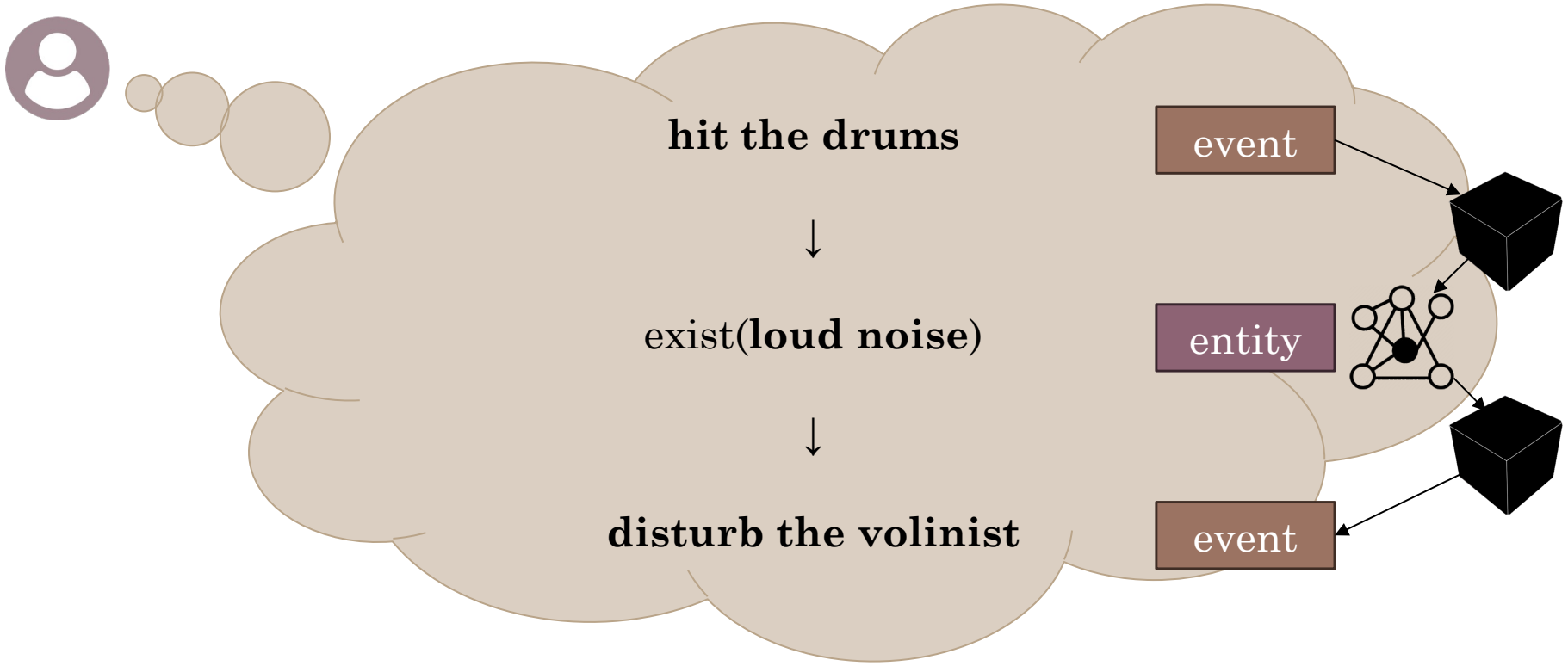
[EACL 2024]

Interim Representation

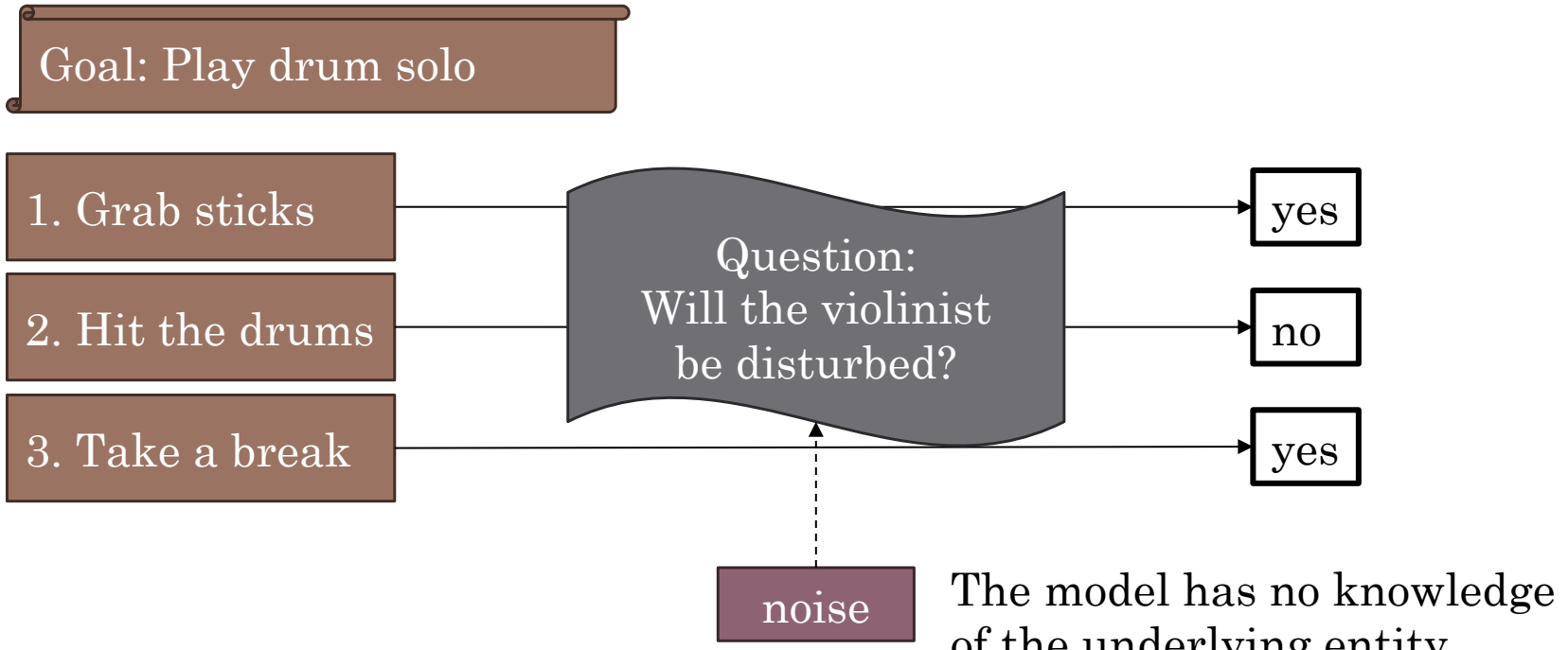


Motivating Example (yet again)

In a music room, a violinist is tuning. A drummer starts playing an elaborate drum solo. What will the violinist say?



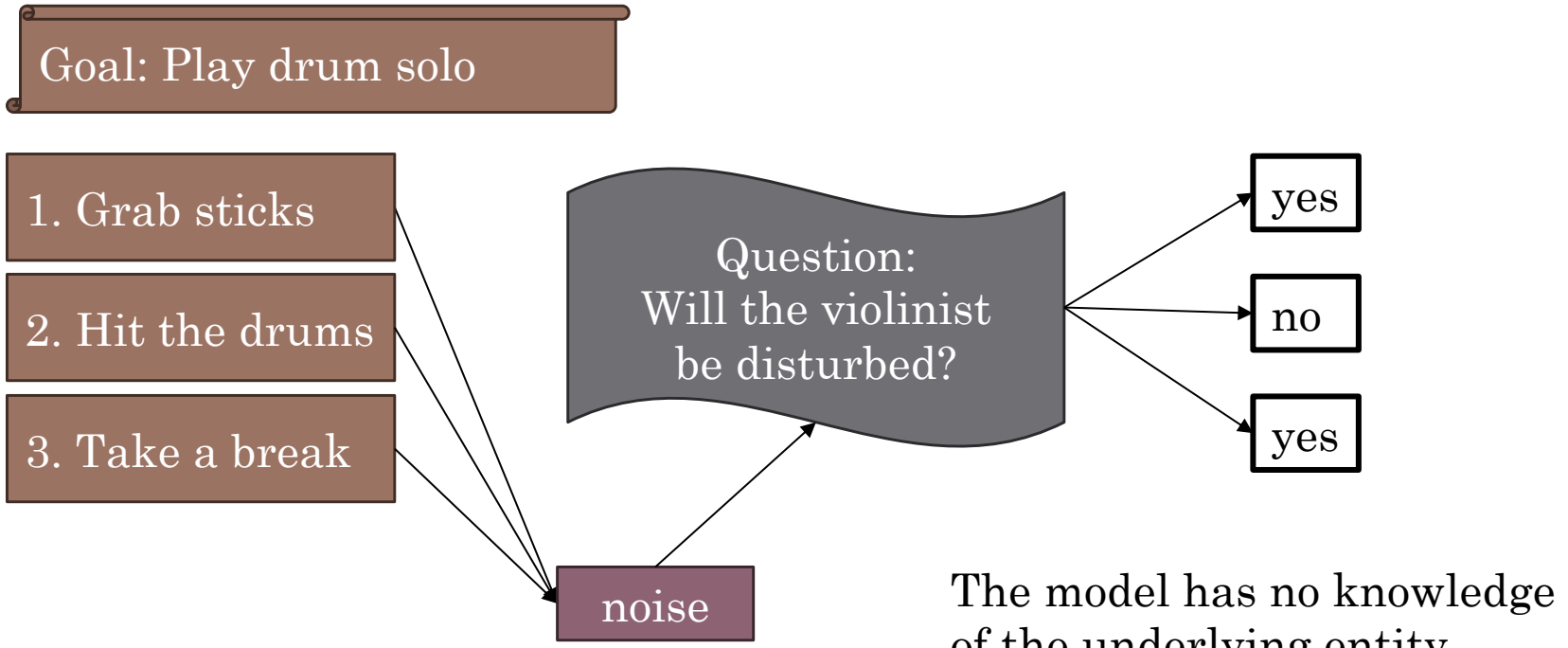
Baseline: End-to-End Approach



Performs no better than random!

[EACL 2023]

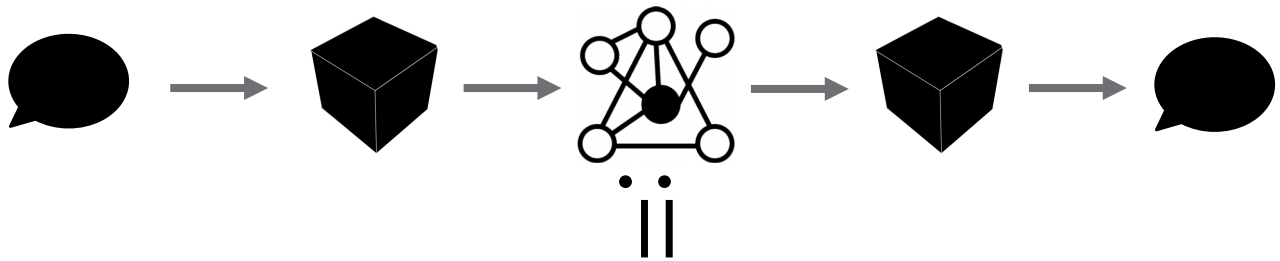
Our Approach: Predict Entity First



The model has no knowledge of the underlying entity.

[EACL 2023]

Entities as Interim Representation



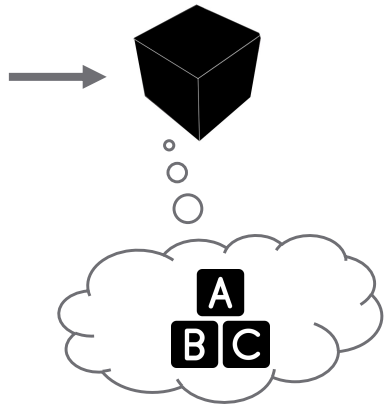
entities	Grab sticks →	Hit drums →	Take break →	…
drummer	idle	idle	playing	idle
noise	none	none	loud	none
…				

- How to feed this to an LLM?

[EACL 2023]

Natural Language Representation

entities	Grab sticks → Hit drums → Take break			
drummer	idle	idle	playing	idle
noise	none	none	loud	none
…				



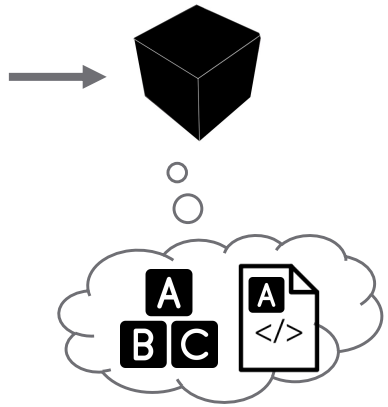
In English

After the step "hit drums", the **drummer** is **playing**, and the **noise** is **loud**.
Given this information, will the violinist be disturbed?

[EACL 2023]

Semi-Symbolic Representation

entities	Grab sticks → Hit drums → Take break			
drummer	idle	idle	playing	idle
noise	none	none	loud	none
…				



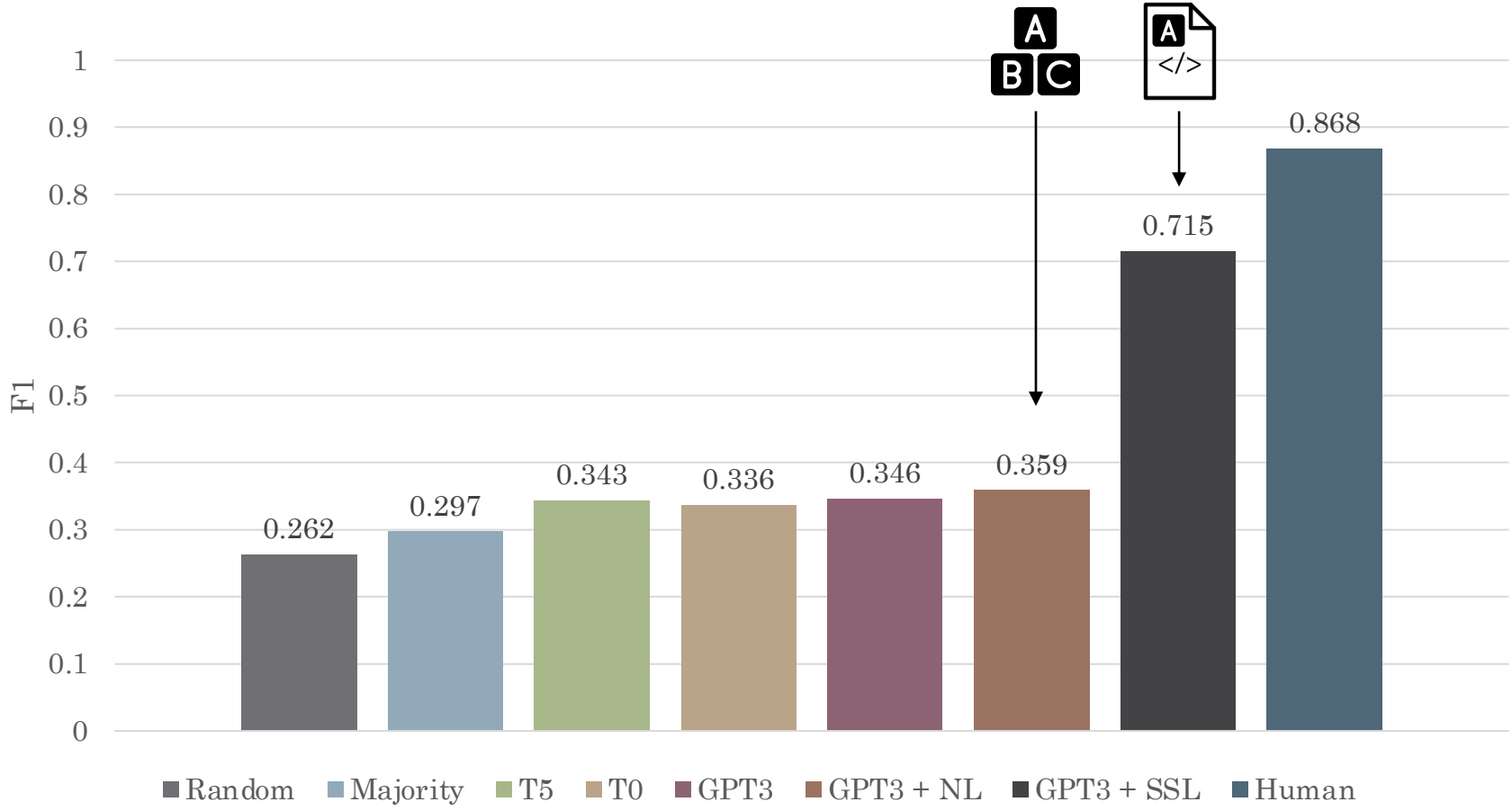
In Python

```
class Play_Drums():
  # Grab sticks, hit drums, take break
  def init(self):
    self.drummer = Drummer(), self.drummer.motion = idle
    self.noise = Noise(), self.noise.level = none
    self.event = “The violinist is disturbed."
  def hit_drums(self):
    self.drummer.motion = playing, self.noise.level = loud
  ...
```

This is not executed!
Just input to LLMs.

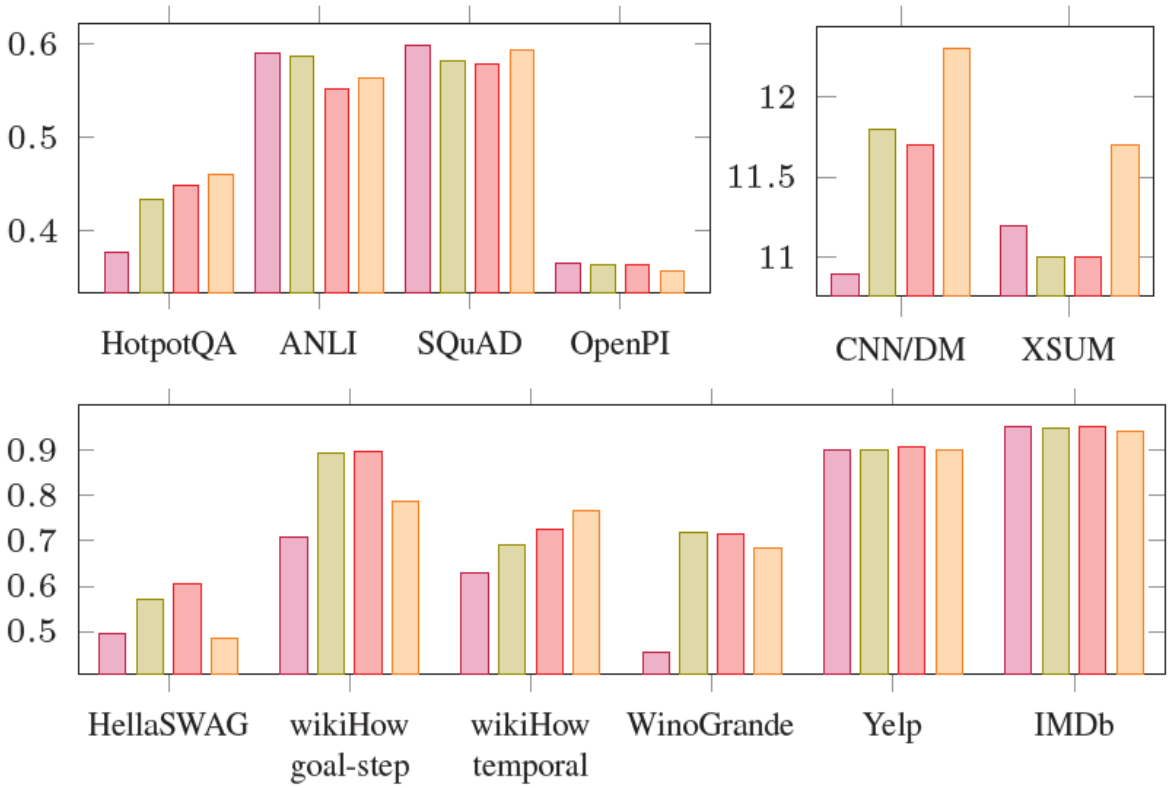
[EACL 2023]

Code-Like Form Works



[EACL 2023]

Code-Like Form *Sometimes* Works



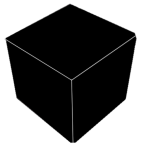
- Many papers suggest code representation works
- Performance of GPT3.5 on a variety of NLP tasks
- Results are mixed

→ Different formulation of code-like prompts

[ACL 2023]

Motivating Example (finally)

In a music room, a violinist is tuning. A drummer starts playing an elaborate drum solo. What will the violinist say?



hit the drums ✓

↓

exist(**loud noise**) ✓

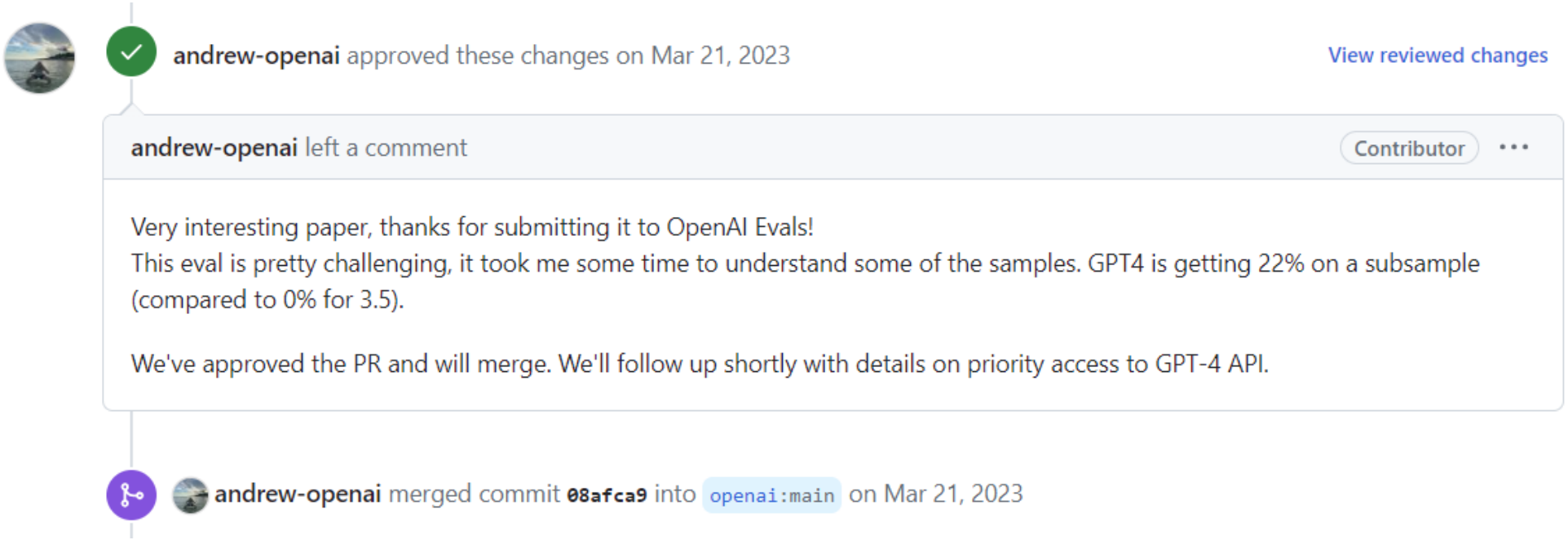
↓

disturb the violinist ✓

"Can you wait until I finish tuning?" ✓

Impact of Our Work

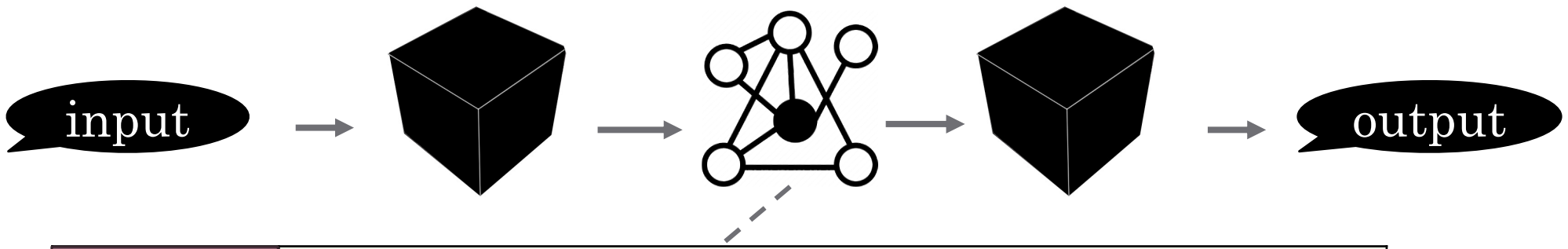
- A part of the evaluation for GPT-4 in [OpenAI Evals]



Summary

- Problem: LLMs cannot effectively reason about the causality among entities and events
- Contribution: an entity-state representation, a semi-symbolic dataset, and prompting LLMs with a code-like structure
- Result: significantly improved performance
- However:

Semi-Symbolic Reasoning



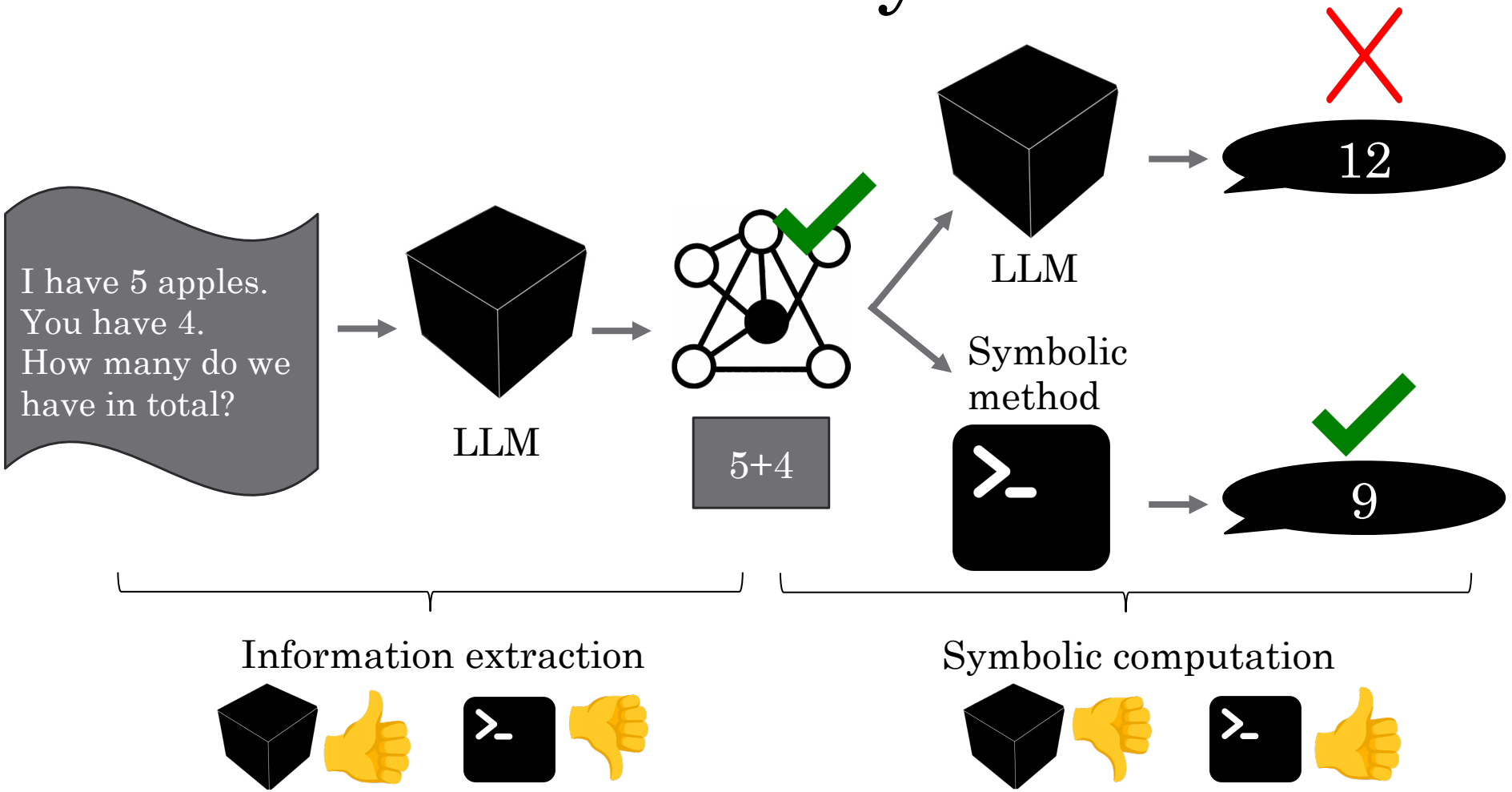
entities	Grab sticks → Hit drums → Take break → …			
drummer	idle	idle	playing	idle
noise	none	none	loud	none
…				

Question:
Will the violinist be disturbed?

:= yes if noise.level = loud
no otherwise

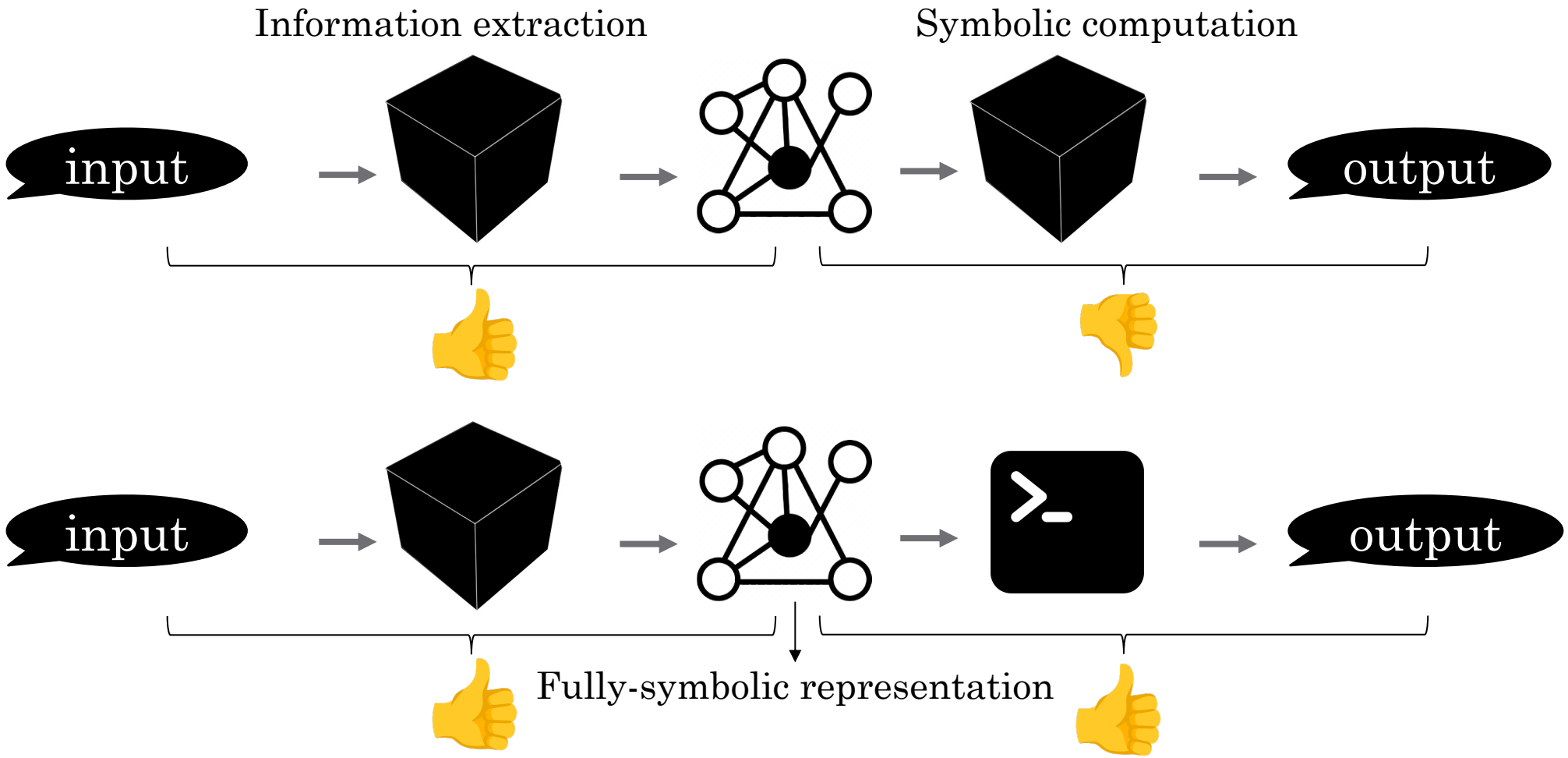
a symbolic formula

LLMs are Bad at Symbols

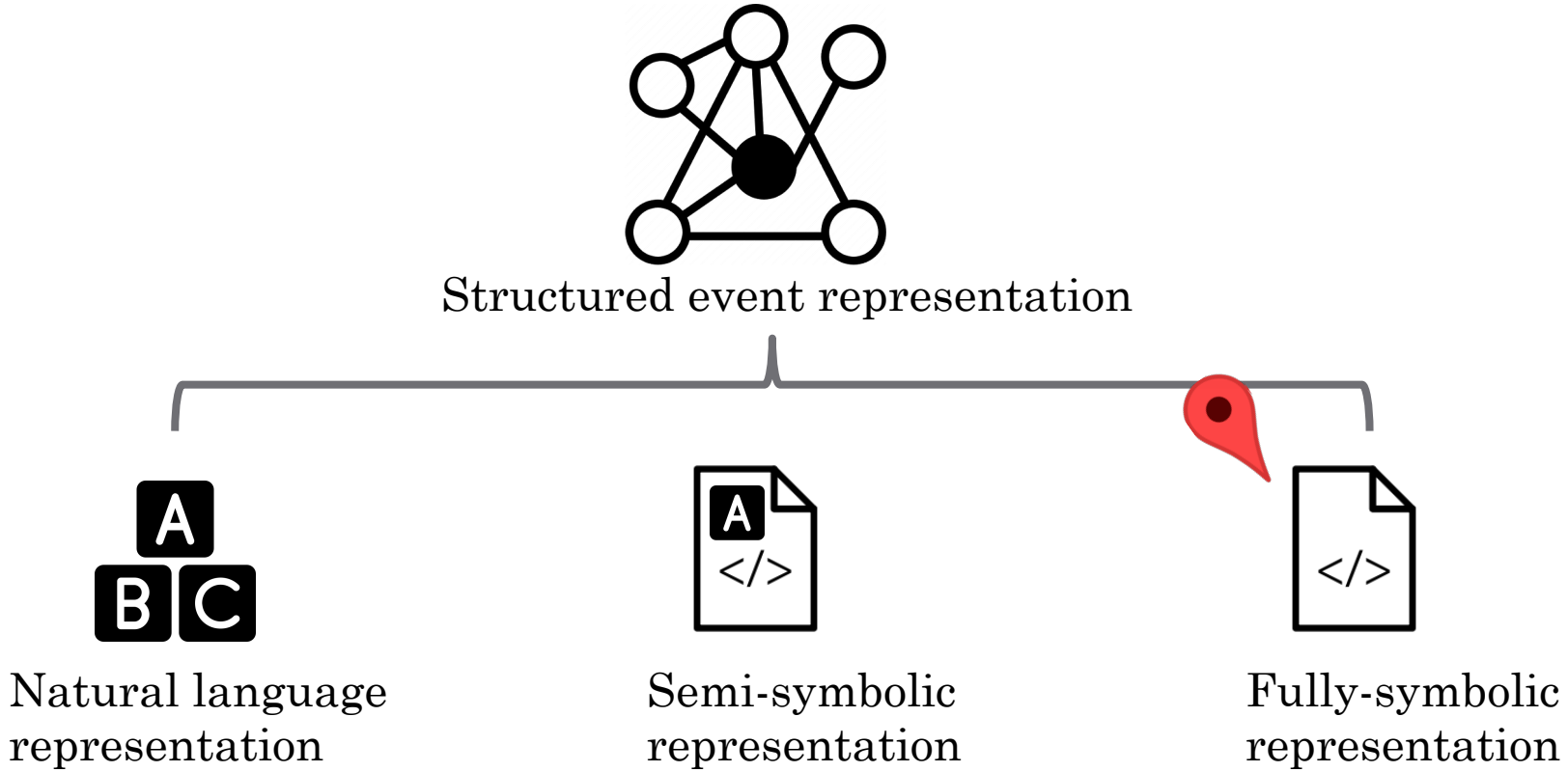


[AACL 2023]

Best of Both Worlds



Roadmap

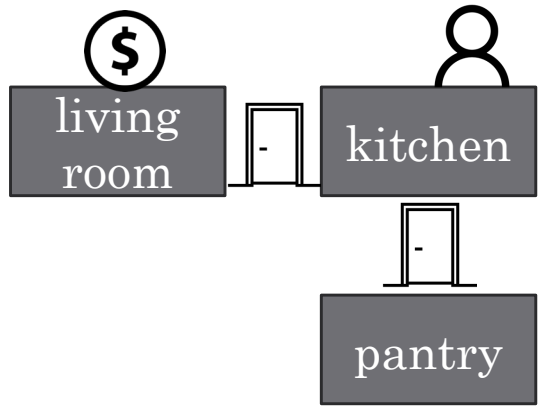

Structured event representation
A
B C
Natural language representation
A
</>
Semi-symbolic representation
</>
Fully-symbolic representation

A Planning Simulation

> You are in the kitchen. You see a counter and some cookware. To the South you see a closed frosted-glass door. To the West you see a closed wooden door.

YOUR GOAL IS TO PICK UP A COIN.

<



[*SEM 2024]

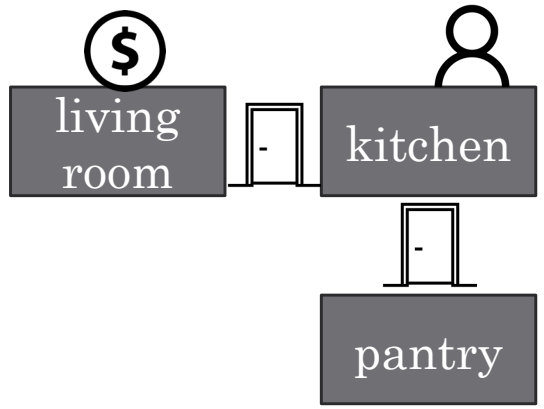
A Planning Simulation (Human)

> You are in the kitchen. You see a counter and some cookware. To the South you see a closed frosted-glass door. To the West you see a closed wooden door.

YOUR GOAL IS TO PICK UP A COIN.

<

I'll try the south first. If I don't find the coin, I'll backtrack and try the west.



[*SEM 2024]

A Planning Simulation (Human)

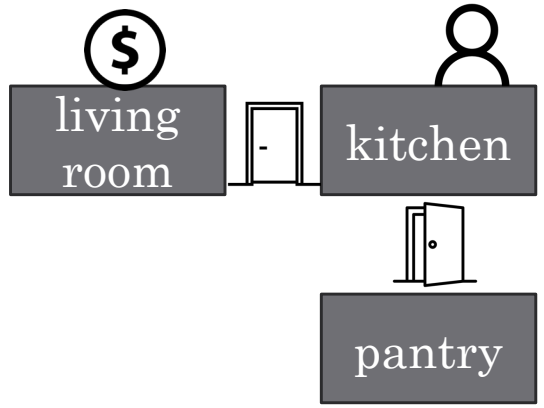
> You are in the kitchen. You see a counter and some cookware. To the South you see a closed frosted-glass door. To the West you see a closed wooden door.

YOUR GOAL IS TO PICK UP A COIN.

< **open door to south**

> You open the door to the South, revealing the pantry.

<



[*SEM 2024]

A Planning Simulation (Human)

> You are in the kitchen. You see a counter and some cookware. To the South you see a closed frosted-glass door. To the West you see a closed wooden door.

YOUR GOAL IS TO PICK UP A COIN.

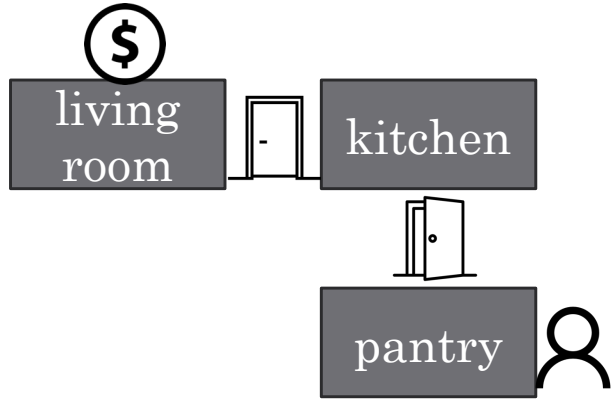
< **open door to south**

> You open the door to the South, revealing the pantry.

< **go south**

> You are in the pantry. You see some snacks. Through an open door, to the North you see the kitchen.

<



[*SEM 2024]

A Planning Simulation (Human)

> You are in the kitchen. You see a counter and some cookware. To the South you see a closed frosted-glass door. To the West you see a closed wooden door.

YOUR GOAL IS TO PICK UP A COIN.

< open door to south

> You open the door to the South, revealing the pantry.

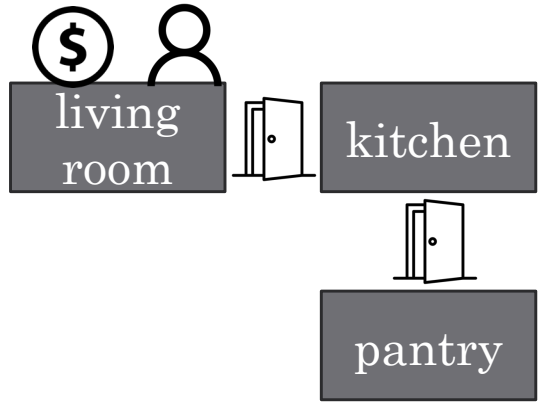
< go south

> You are in the pantry. You see [some irrelevant objects]. Through an open door, to the North you see the kitchen.

< go north … < open door to west … < go west

> You are in the living room. YOU SEE A COIN. Through an open door, to the East you see the kitchen.

< pick up coin

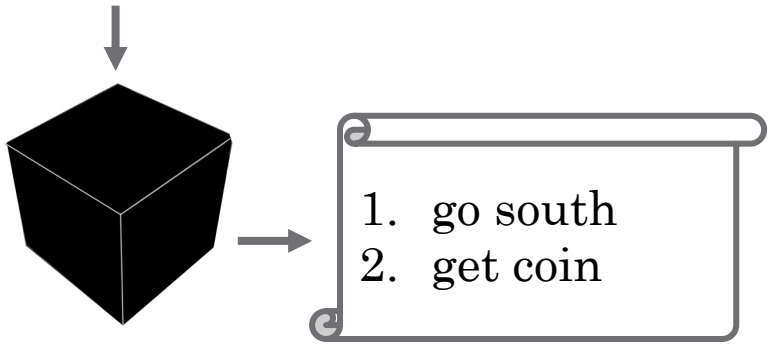


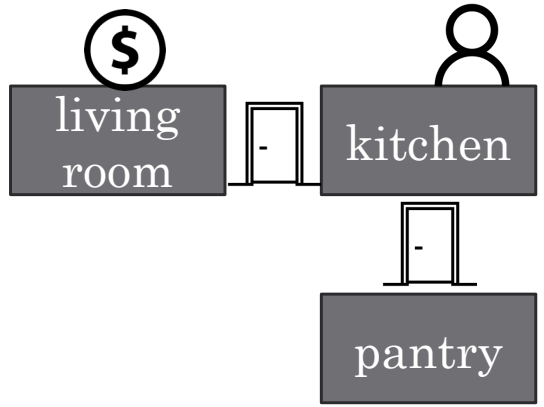
[*SEM 2024]

A Planning Simulation (LLM)

> You are in the kitchen. You see a counter and some cookware. To the South you see a closed frosted-glass door. To the West you see a closed wooden door.

YOUR GOAL IS TO PICK UP A COIN.





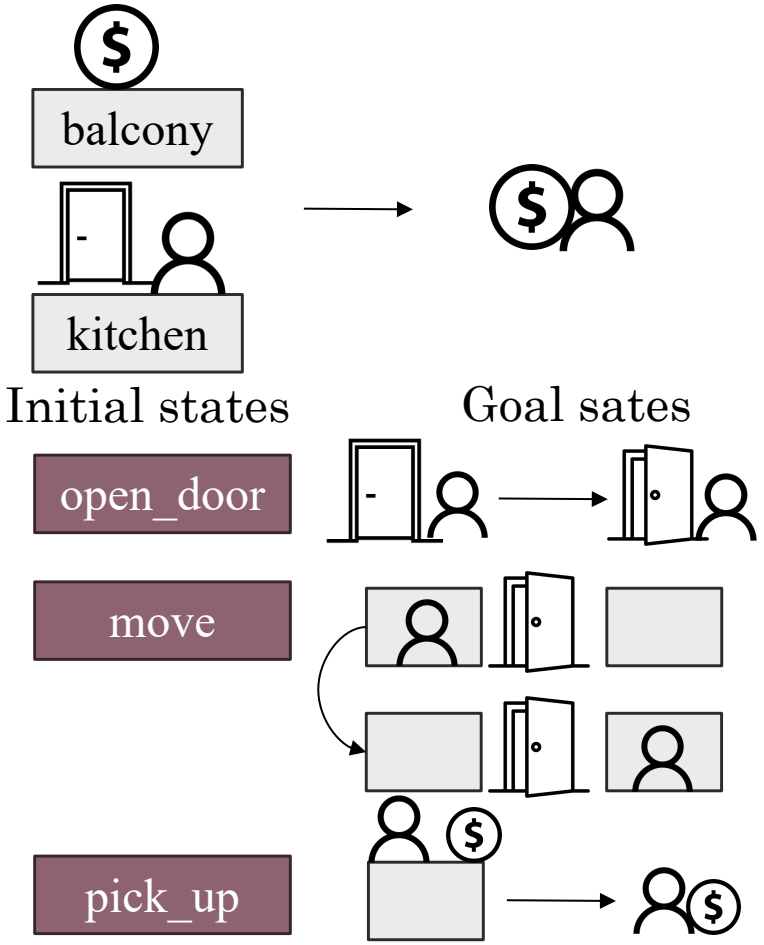
< go south

Door not open

LLMs are not good at symbolic planning

[*SEM 2024]

Classical Symbolic Planning



Problem File (Entity States)

Initial states:

```
at(you, kitchen)
at(coin, balcony)
closed_door(kitchen, balcony)
```

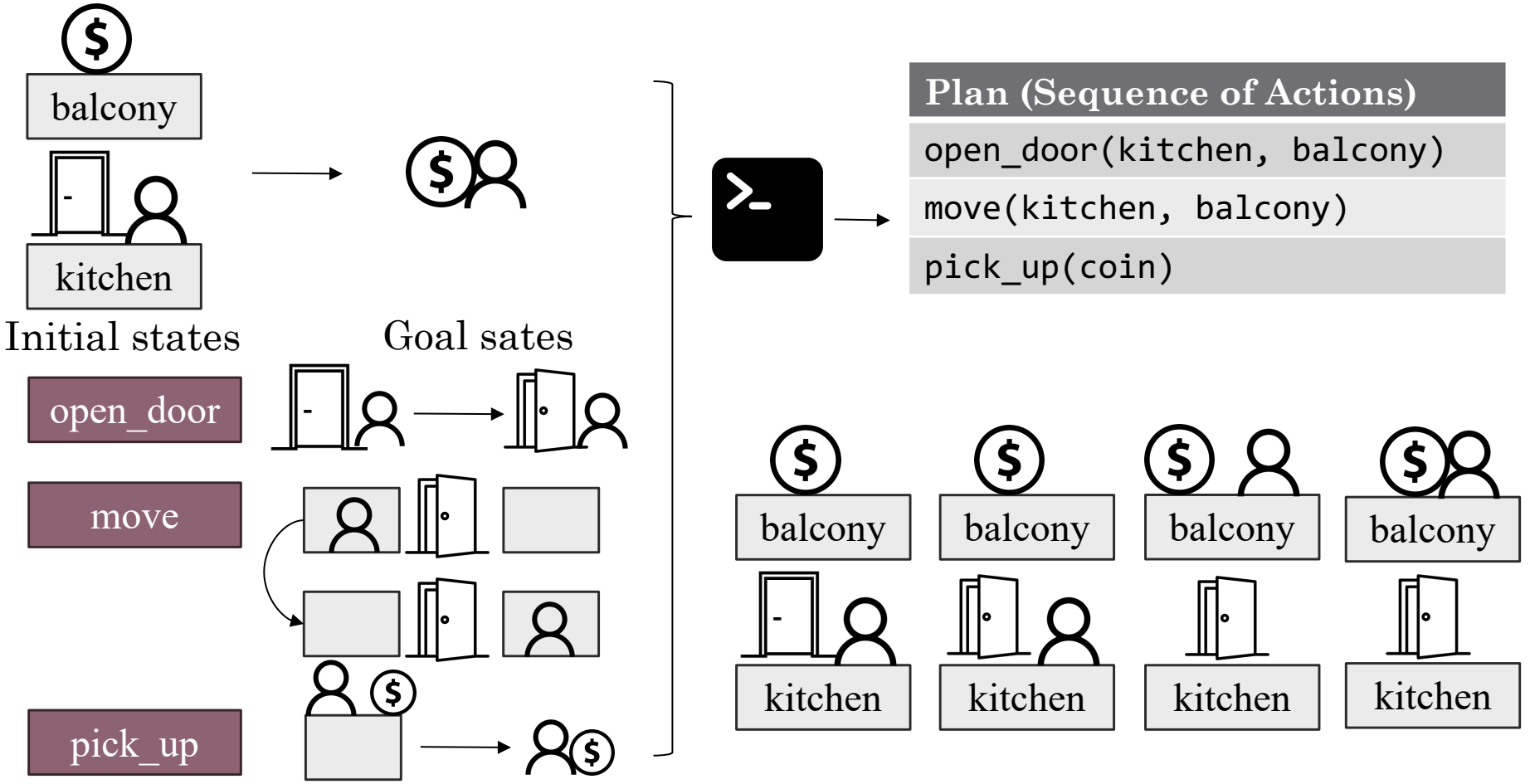
Goal states:

```
have(you, coin)
```

Domain File (Action Models)

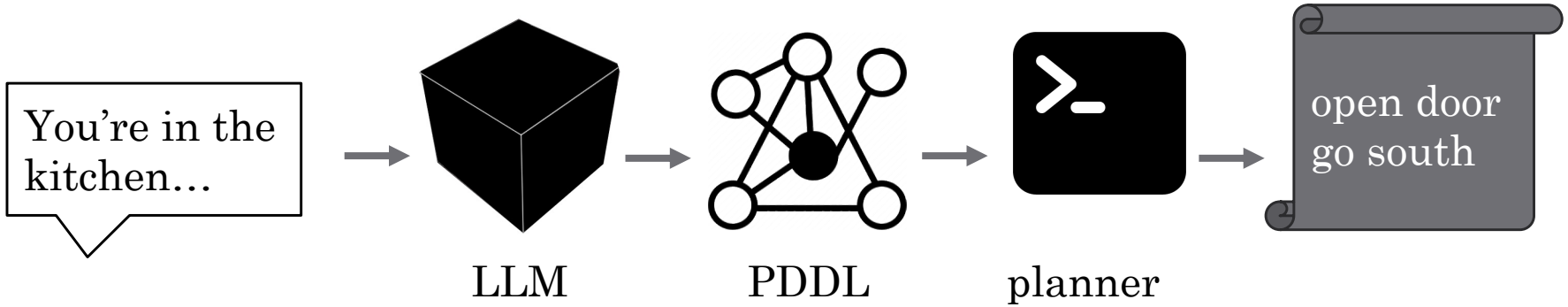
Action	Precondition	Effect
`open_door`	`at(you, r1), closed_door(r1, r2)`	`open_door(r1, r2)`
`move`	`at(you, r1), open_door(r1, r2)`	`at(you, r2)`
`pick_up`	`at(you, r1), at(obj, r1)`	`have(you, obj)`

Classical Symbolic Planning



Best of Both Worlds

- LLMs translate input to a symbolic **world model**



Problem File (Entity States)

Initial states:
`at(you, kitchen), at(coin, balcony)`
`closed_door(kitchen, balcony)`

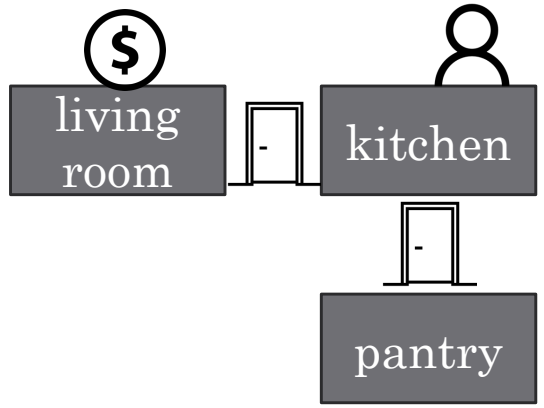
Goal states: `have(you, coin)`

Domain File (Action Models)		
Action	Precondition	Effect
`open_door`	`at(you, r1), closed_door(r1, r2)`	`open_door(r1, r2)`
`move`	`at(you, r1), open_door(r1, r2)`	`at(you, r2)`

A Planning Simulation (LLM+)

> You are in the kitchen. You see a counter and some cookware. To the South you see a closed frosted-glass door. To the West you see a closed wooden door.

YOUR GOAL IS TO PICK UP A COIN.



<

Problem File

Initial states:

```
at(you, kitchen)
visited(kitchen)
closed_door(kitchen, l1, south)
```

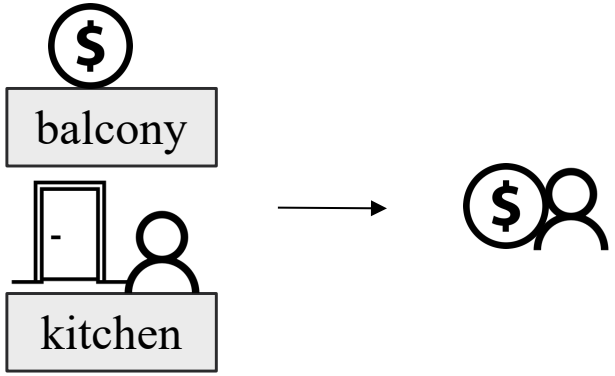
Goal states:

```
have(you, coin)
```

Fails: coin is unobserved!

[*SEM 2024]

Fully vs. Partially Observed



Problem File

Initial states:

```
at(you, kitchen)
at(coin, balcony)
closed_door(kitchen, balcony)
```

Goal states:

```
have(you, coin)
```

- We have assumed a fully-observed environment
- Challenge: we don't know where the coin is, or how the rooms are connected

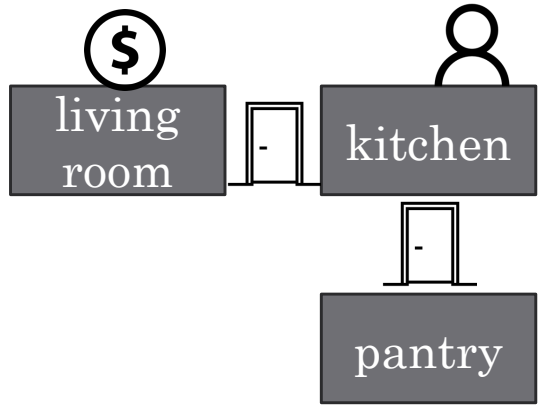
[*SEM 2024]

A Planning Simulation (LLM+)

> You are in the kitchen. You see a counter and some cookware. To the South you see a closed frosted-glass door. To the West you see a closed wooden door.

YOUR GOAL IS TO PICK UP A COIN.

< Idea: **decompose** the goal into a sub-goal



Problem File

Initial states:
`at(you, kitchen)`
`visited(kitchen)`
`closed_door(kitchen, l1, south)`

Goal states: `∃loc, at(you, loc) & ¬visited(loc)`

Now, progress can be made.

Note: this is a sub-goal to go to an unvisited location.

A Planning Simulation (LLM+)

Predicted PF

Problem File

Initial states:
`at(you, kitchen)`
`Visited(kitchen)`
`closed_door(kitchen, l1, south)`

Goal states:
`∃loc, at(you, loc) & ¬visited(loc)`

Provided DF

Domain File

Action	Precondition	Effect
`open_door`	`at(you, r1), closed_door(r1, r2)`	`open_door(r1, r2)`
`move`	`at(you, r1), open_door(r1, r2)`	`at(you, r2)`

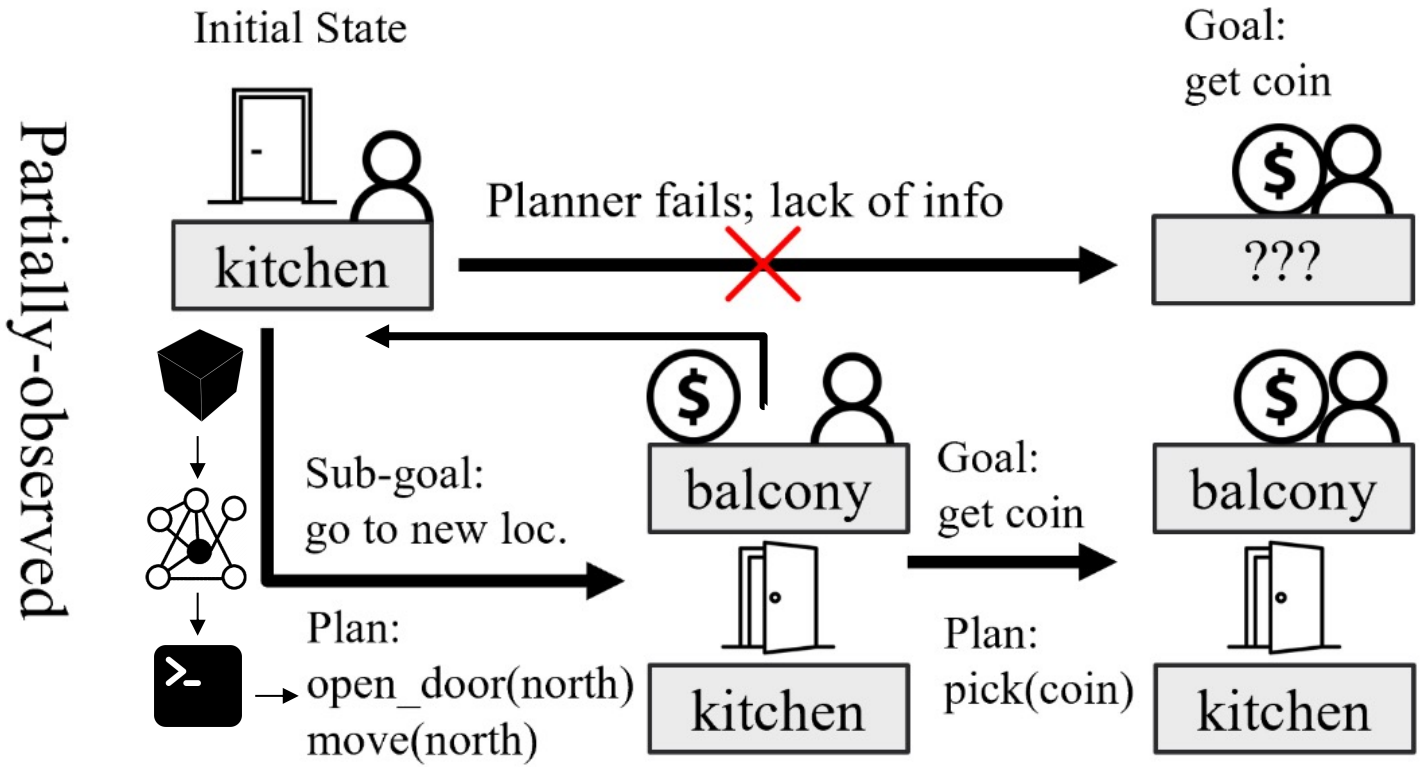
Plan

`open_door(south)`
`go(south)`

If predicted PF is correct, the plan *must* be correct.

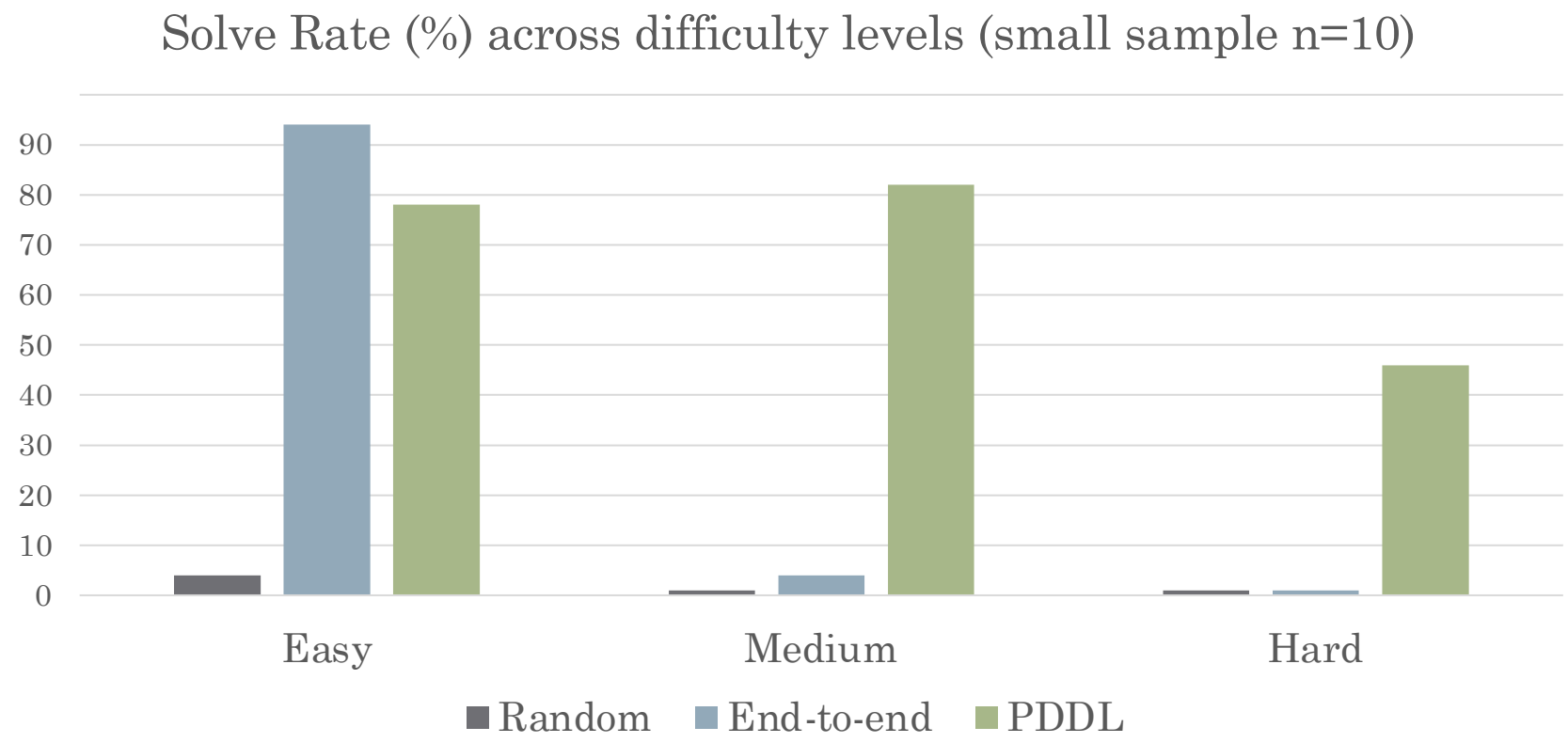
[*SEM 2024]

Planning during Exploration



[*SEM 2024]

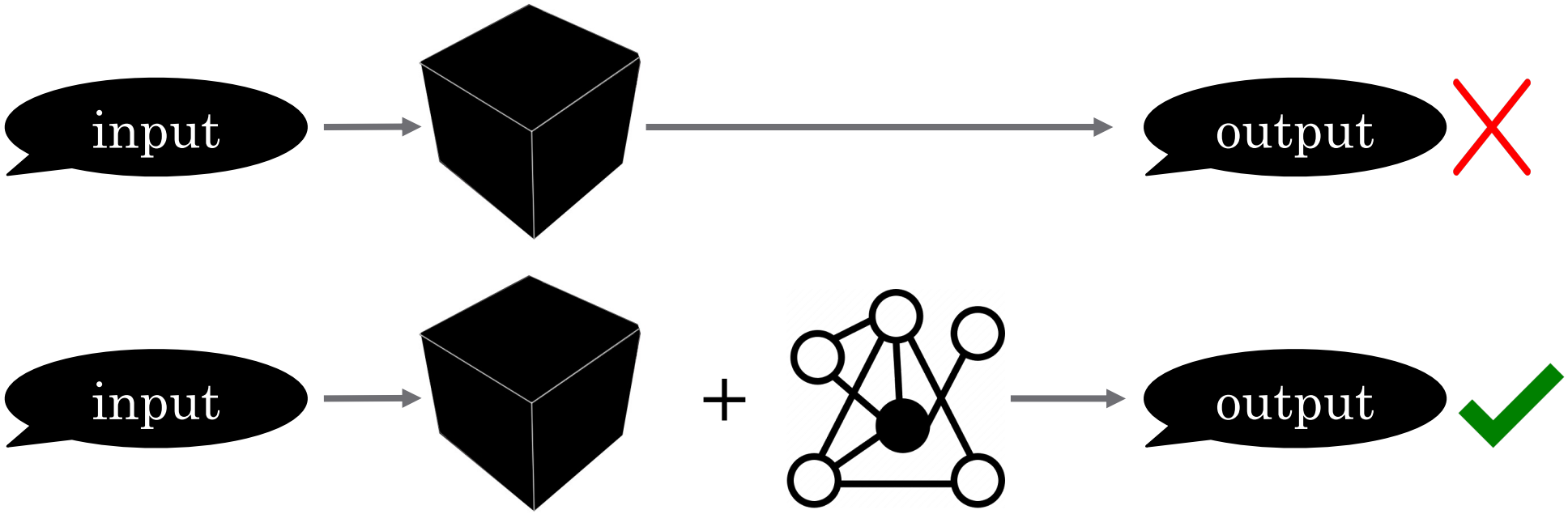
LLMs are Better at World-Modeling than Planning

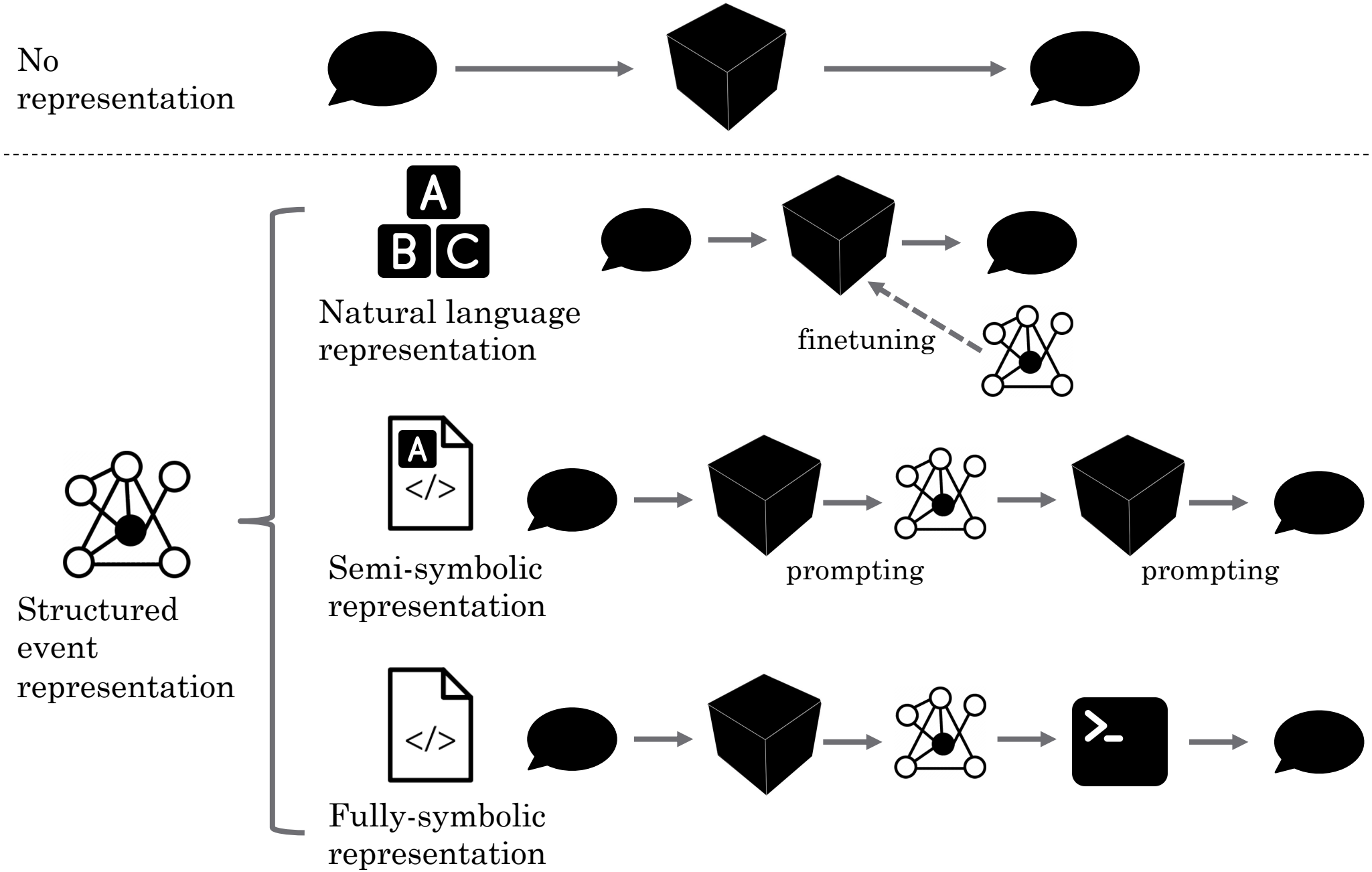


[*SEM 2024]

Summary

- End-to-end LLM is not the way to go for event reasoning
- LLM should work with a **structured** representation

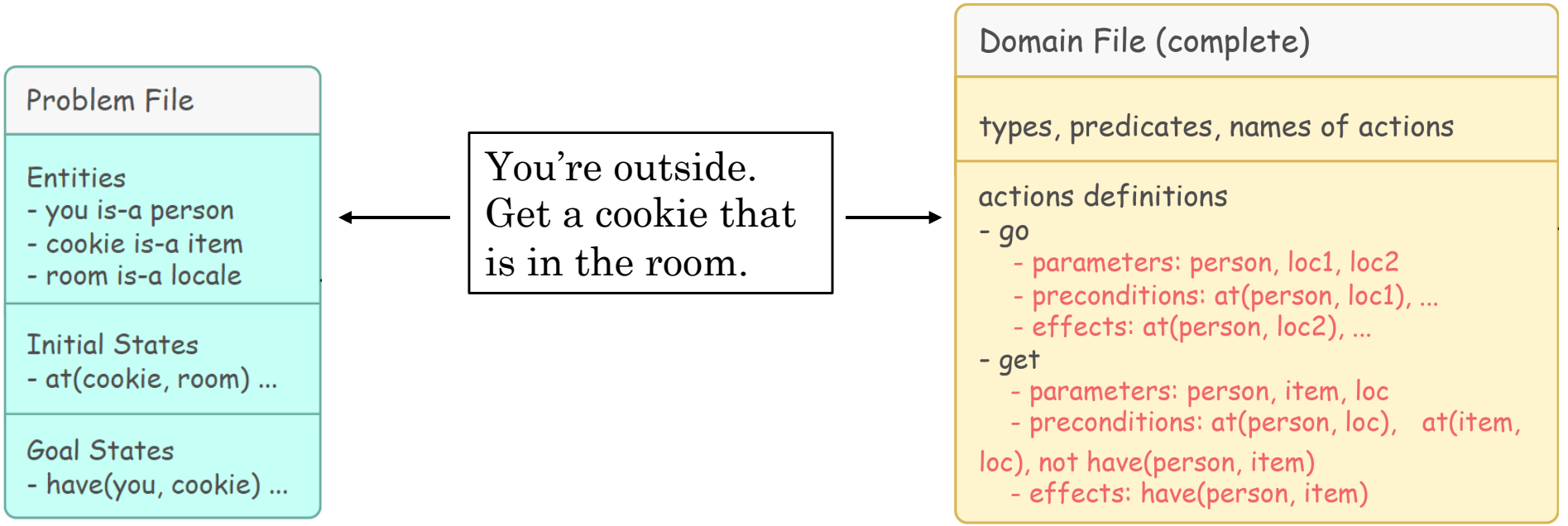



No representation
Natural language representation
finetuning
Structured event representation
Semi-symbolic representation
prompting
prompting
Fully-symbolic representation

Future Work

Short-Term

- Previously, we only modeled entity states (problem file)
- Also model **actions** (domain file)



Short-Term

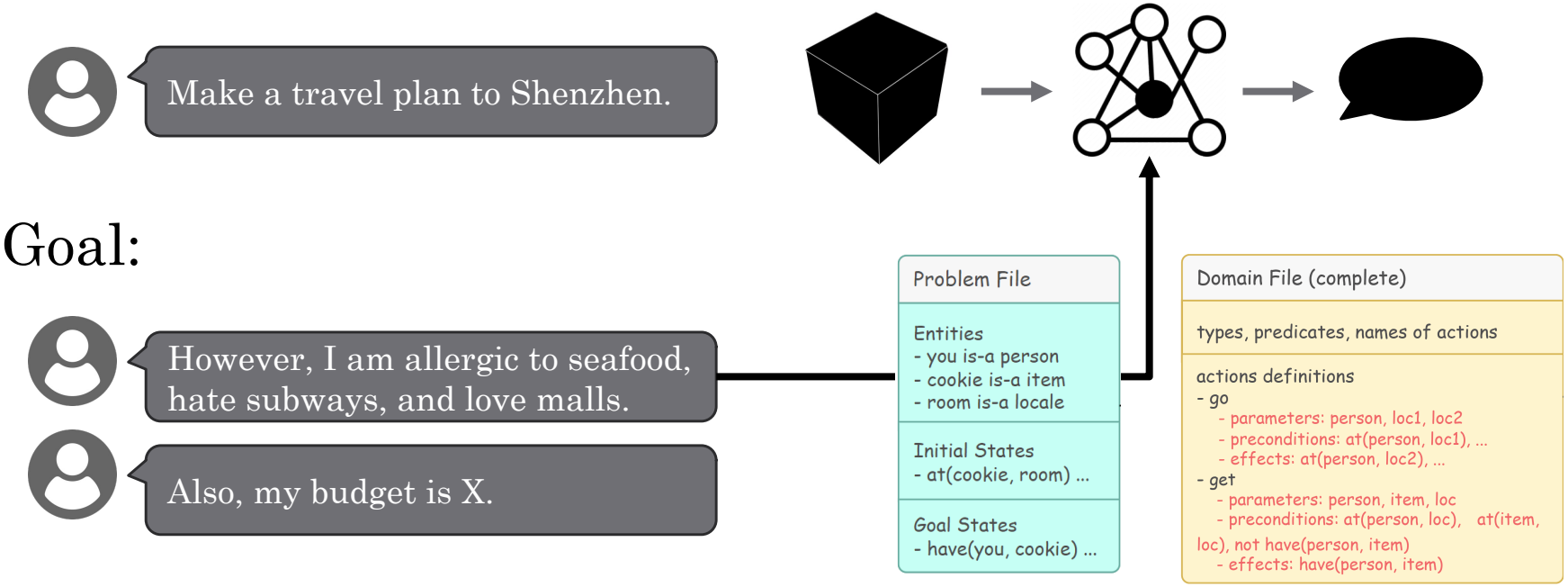
- Previously, we only modeled entity states (problem file)
- Also model **actions** (domain file)
- The representation iteratively grows and refine

(:action Charge
Precondition: cord plugged in
Effect: phone starts charging)

(:action Charge
Precondion: cord plugged in, **outlet on**
Effect: phone starts charging)

Long-Term

- Build models that can dynamically customize to *everyone*'s circumstances, constraints, and requirements
- Previously:



- Goal:

Belief in Neurosymbolic Methods

Purely data-driven methods are bound by what has been represented in the data

- Has always succumbed to **long-tail problem**
- Unlikely to solve problems for underrepresented individuals

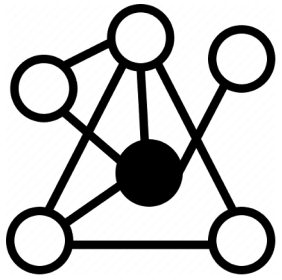
LLM + symbols is highly promising because:

- LLM can flexibly generate and generalize symbols
- Symbols can lead to precise and interpretable reasoning

Acknowledgements

- Dedicated to Drago.
- Thank you, Chris!
- Thank you, Rada!
- Thank you, mom and dad!
- Thank you, Vivienne!
- Thank you, Niket, Veronica, and Shuyan!
- Thank you, Pete, Peter, Steve, Rui, Graham, Dan, Marianna, and Mark!
- Thank you, Ziyang and Jiani!
- Thank you, Jeffrey, Hainiu, Joey, and Tianyi!
- Thank you, Ellie, Reno, Daphne, Jie, Lara, Aditya, Artemis, Yue, Bryan, Alyssa, Ajay, Liam, and Samar!

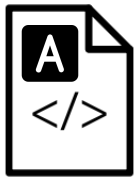
Structured Event Reasoning with Large Language Models



Structured event representation

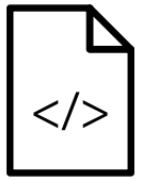
Natural language representation

EMNLP 2020, TMLR, AACL 2020, INLG 2021



Semi-symbolic representation

EACL 2023, ACL 2023, EACL 2024



Fully-symbolic representation

*SEM 2024

Other publications:

*SEM 2019
Semantic Similarity
EMNLP 2020
Text Simplification
EMNLP 2021
Visual Event Reasoning
ACL 2022
Hierarchical Procedures
NAACL 2022
Semantic Role Labeling
NAACL 2022
Recursive Noun Phrases
EMNLP 2022
Entity Linking
AAAI 2023
Music Generation
AACL 2023
Faithful Chain-of-Thought
ACL 2023
Tool for Schema Induction
ACL 2023 workshop
Prompting LLMs w/ code

Structured Event Reasoning with Large Language Models

Li “Harry” Zhang

University of Pennsylvania (Ph.D., 2019-2024)