ReasoningFlow:
Semantic Structures of Complex Reasoning Traces

Jinu Lee, Sagnik Mukherjee, Dilek Hakkani-Tur, Julia Hockenmaier (UIUC)

ArgMining Workshop @ ACL 2025

Presented at: EleutherAI, 1/21/2026

About me

Jinu Lee

Research interests

Evaluation of reasoning traces, Natural language formalization, Improving LLM reasoning

Education

Ph.D. ***Comp. Sci.*** University of Illinois Urbana-Champaign *Aug 2024 - Now*
(Advised by: Dr. Julia Hockenmaier)

B.S. ***Comp. Sci. Eng.*** Seoul National University *Mar 2018 - Aug 2024*
(Worked with: Dr. Seungwon Hwang and Dr. Wonseok Hwang)

Careers

Research Intern Microsoft Research *May 2025 - Aug 2025*

Research Engineer LBOX *(legal tech startup)* *Jul 2023 - Jun 2024*

Researcher NCSOFT Language AI Lab *Jun 2020 - Nov 2020*

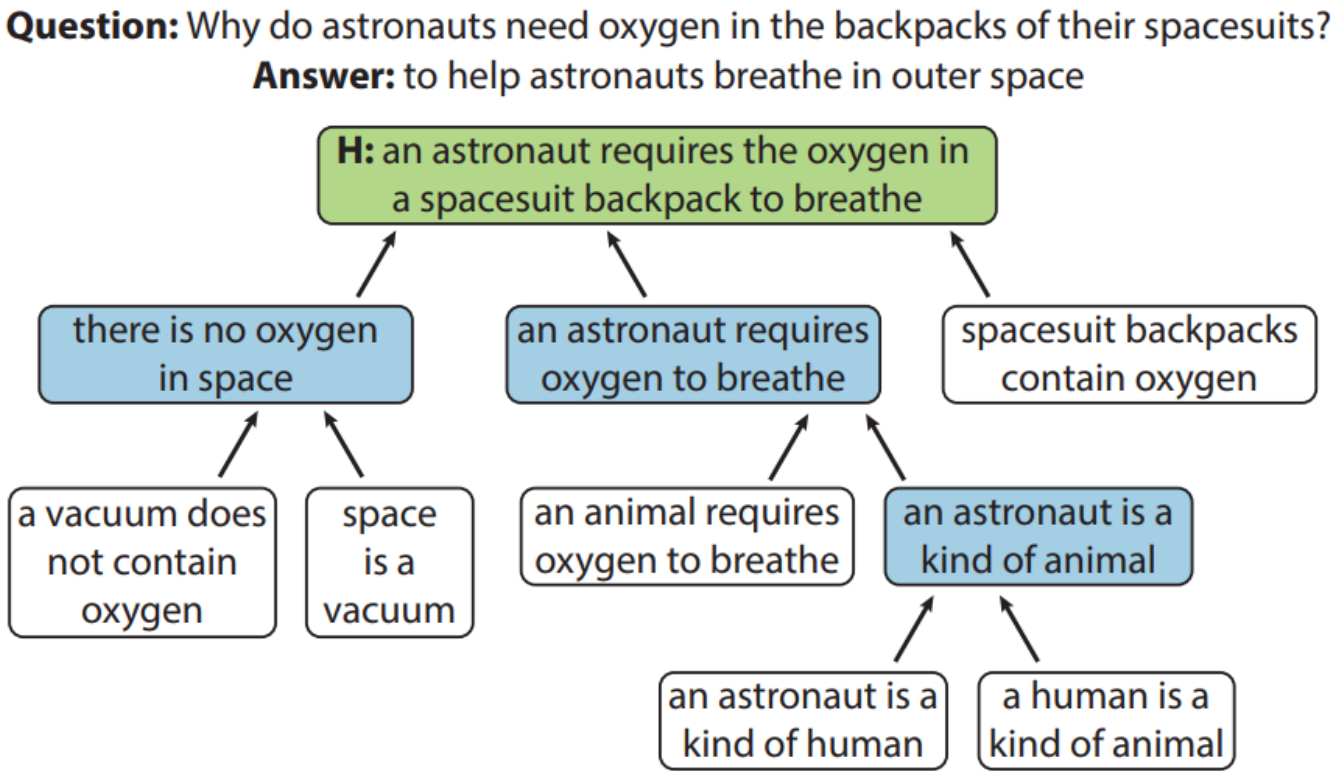
Contents

Introduction

Q. Why is structure important in reasoning?

A. Because reasoning is a structured process.

- Reasoning was commonly expressed as **trees** across multiple disciplines (logic, law, argumentation, …)
- Conclusions are deduced from their premises, in a recursive manner



EntailmentBank
(Dalvi et al., 2021)

Introduction

Modern Chain-of-thought traces are more than an entailment tree.

DeepSeek-R1-style traces are **non-linear and verbose**

- Includes non-propositional elements like planning and reflection
- These elements form diverse semantic behaviors, which cannot be explained by entailment.

Self-reflection
This seems correct.

Verifications
Let me check my answer.

Backtracking
Let's try a different approach.

Top-down planning
First, we need to calculate...

Introduction

Structures of long text: **Argumentation parsing / Discourse parsing**

- Divide long text into small units (Premises+Claims / Elementary Discourse Units)
- Identify labeled directed edges between these units

Q. Can we apply existing datasets/schemes directly to LLM reasoning traces?

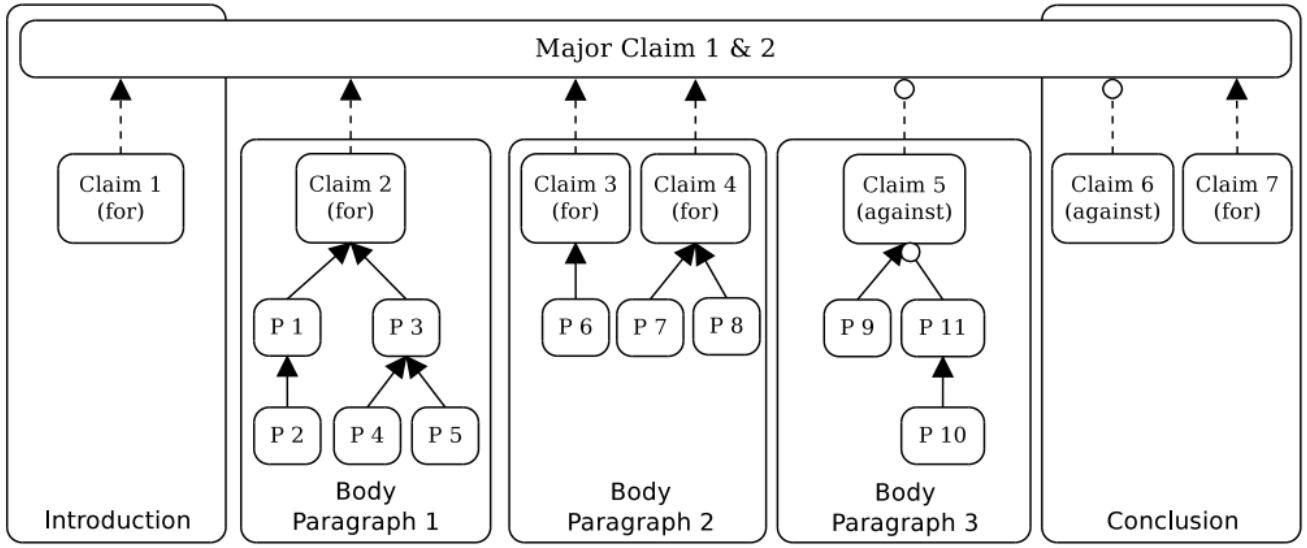
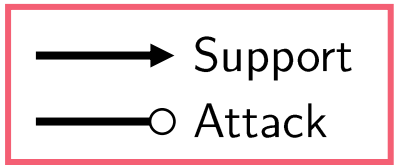


Figure 2
Argumentation structure of the example essay. Arrows indicate argumentative relations. Arrowheads denote argumentative support relations and circleheads attack relations. Dashed lines indicate relations that are encoded in the stance attributes of claims. "P" denotes premises.



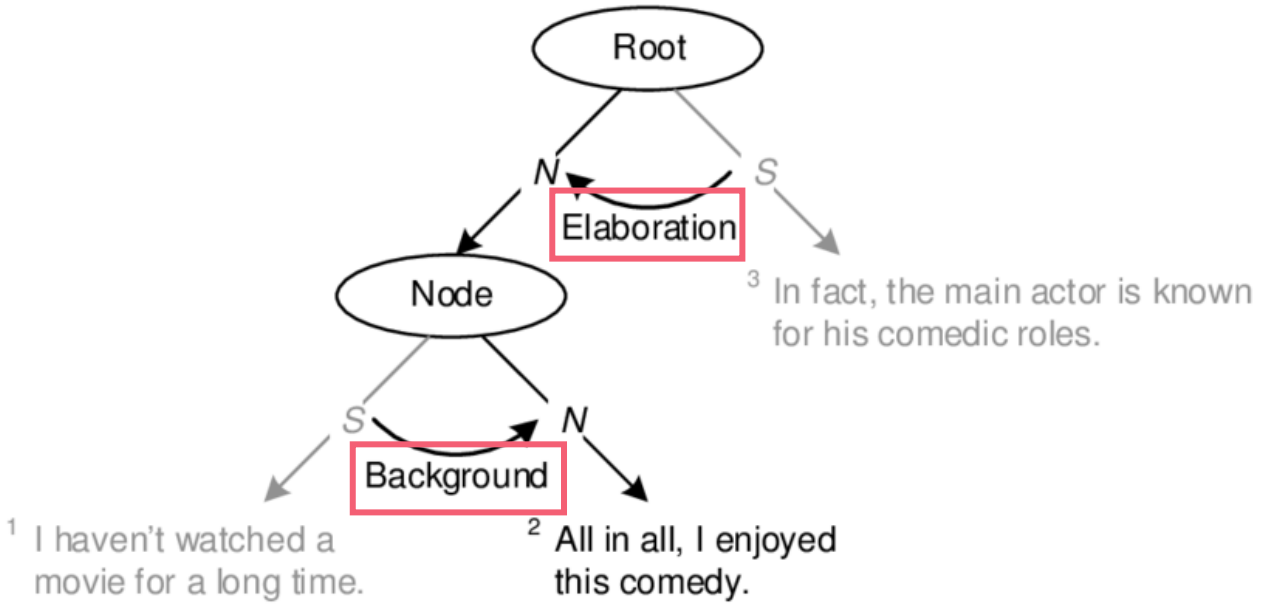
Parsing argumentation structures in persuasive ssays (Stab & Gurevych, 2017)

Introduction

Structures of long text: **Argumentation parsing / Discourse parsing**

- Divide long text into small units (Premises+Claims / Elementary Discourse Units)
- Identify labeled directed edges between these units

Q. Can we apply existing datasets/schemes directly to LLM reasoning traces?



Example of a discourse tree (RST) (Staus, 2017)

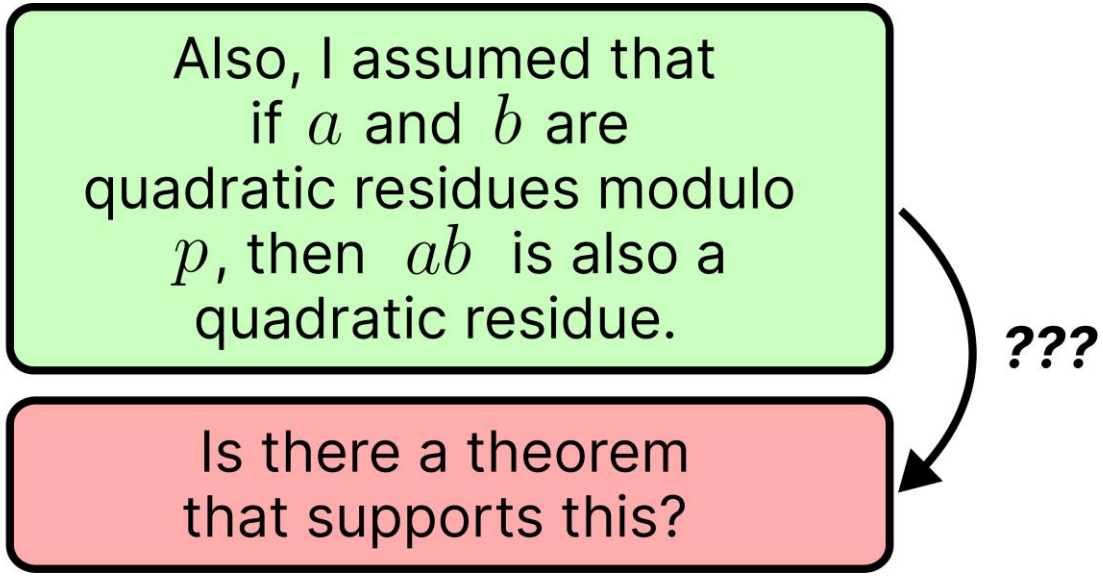
Introduction

Structures of long text: **Argumentation parsing / Discourse parsing**

Problem? These labels do not directly correspond with LLM reasoning

For instance, no existing theories in argumentation/discourse covers this relation:

- The following step **plans to verify** the previous step.



ReasoningFlow labels

Design principles:

1. **Left-to-right edges = Directed acyclic graphs**
   - Premises always precede conclusions
   - Reflects the autoregressive nature of left-ro-right LLMs
   - *cf.* Human argumentation/discourse theory allows both directions (forward/backward)
2. **Data-driven labels**
   - No existing theory fully covers the semantic structures of LLM-generated reasoning trace
   - Develop a new label set tailored to LLM reasoning traces

Authors have manually analyzed reasoning traces and developed the tagset.

ReasoningFlow labels

Node labels: 8 semantic roles of reasoning units (< sentence)

- Reasoning: Deriving a new fact by deduction/induction, ...
- Reflection: Assessing other units (confidence/feelings, ...)
- Planning: Introducing the contents of upcoming units
- Fact: General external knowledge
- (4 more labels)

trace17: So, sqrt(1.6/0.4) = sqrt(4) = 2.

trace18: Therefore, f' = 500 Hz * 2 = 1000 Hz.

trace19: Wait, is that correct?

trace20: I feel like there might be more to it because of relativity.

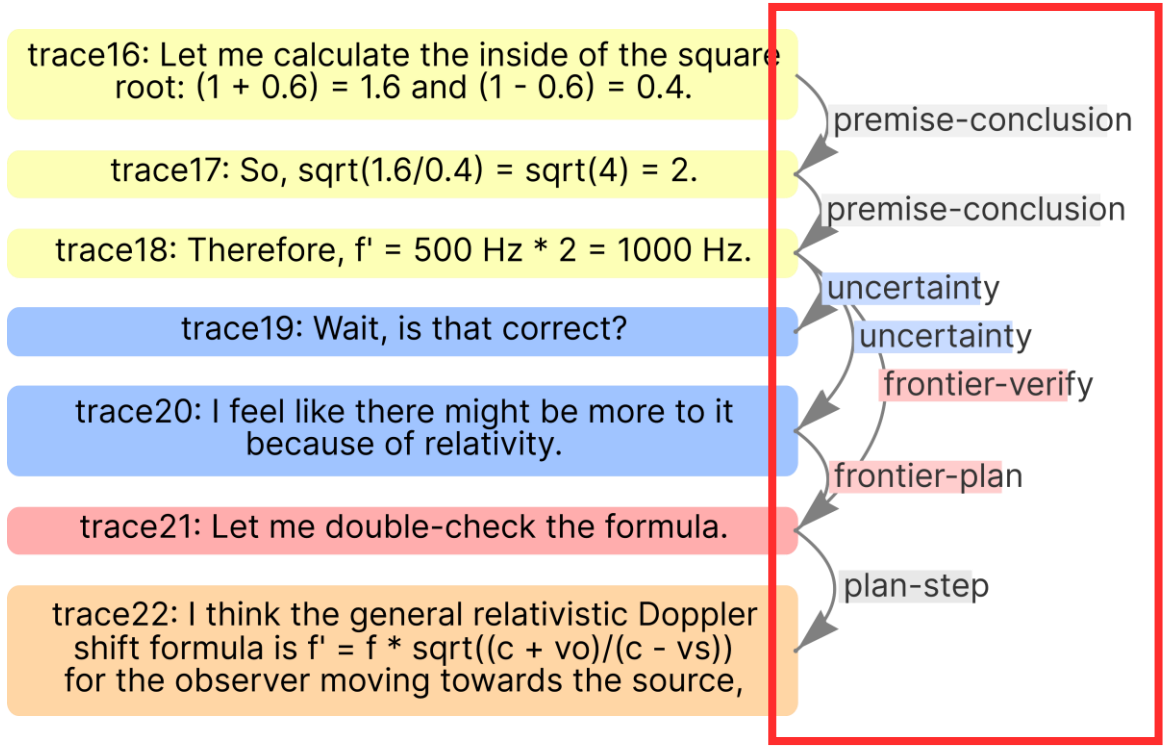
trace21: Let me double-check the formula.

trace22: I think the general relativistic Doppler shift formula is f' = f * sqrt((c + vo)/(c - vs)) for the observer moving towards the source,

ReasoningFlow labels

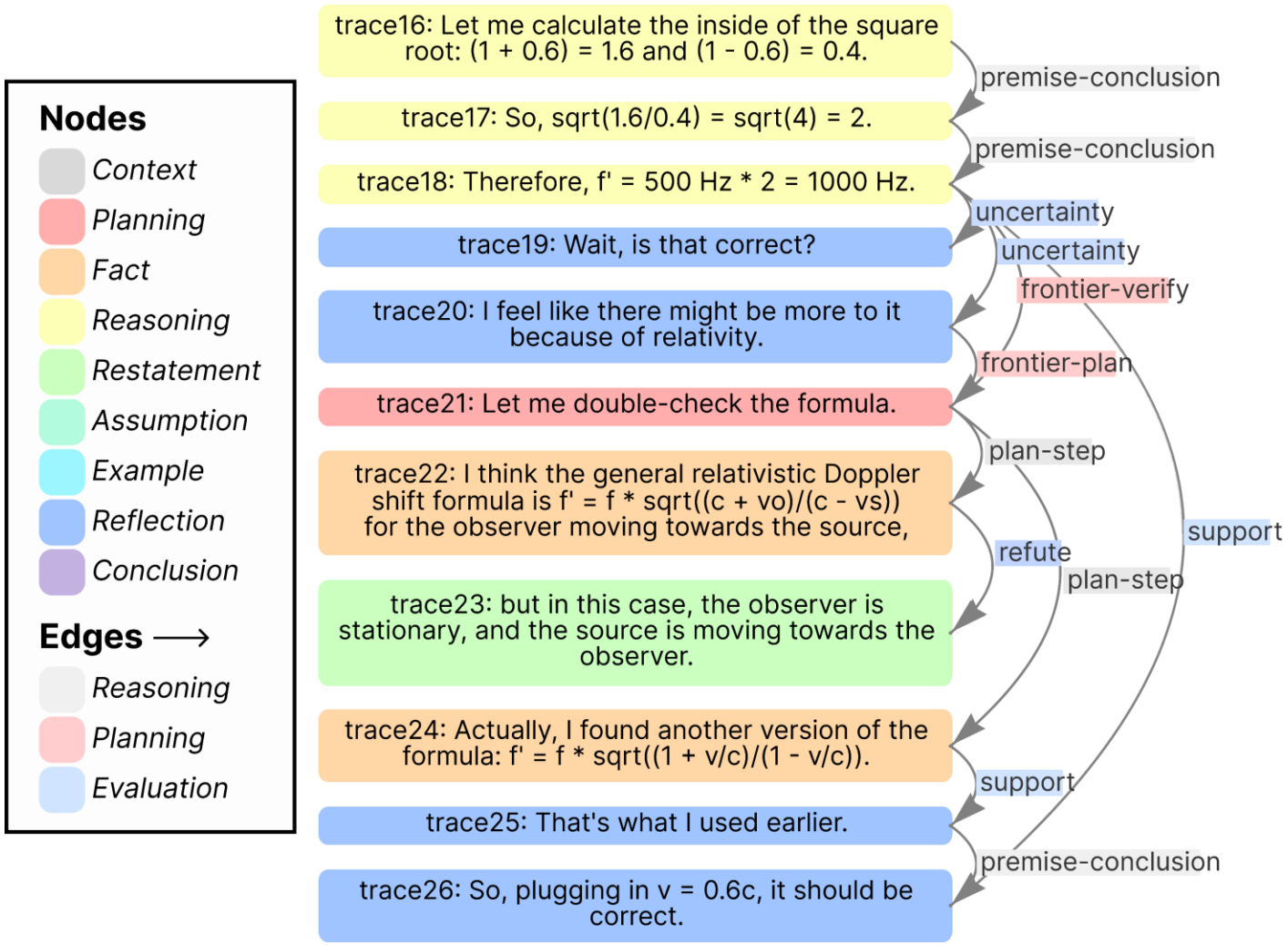
Edge labels: 14 relation types under three major categories (Reasoning, Planning, Evaluation)

- Reasoning: Deriving a **new fact** based on previous units
  - *premise-conclusion* (deduction), *plan-step*, ...
- Planning: Deriving **plans for future steps** from previous units
  - *frontier-plan* (planning based on current knowledge), *frontier-verify* (verification), ...
- Evaluation: Evaluating previous units
  - *support*, *refute*, *uncertainty*



ReasoningFlow labels

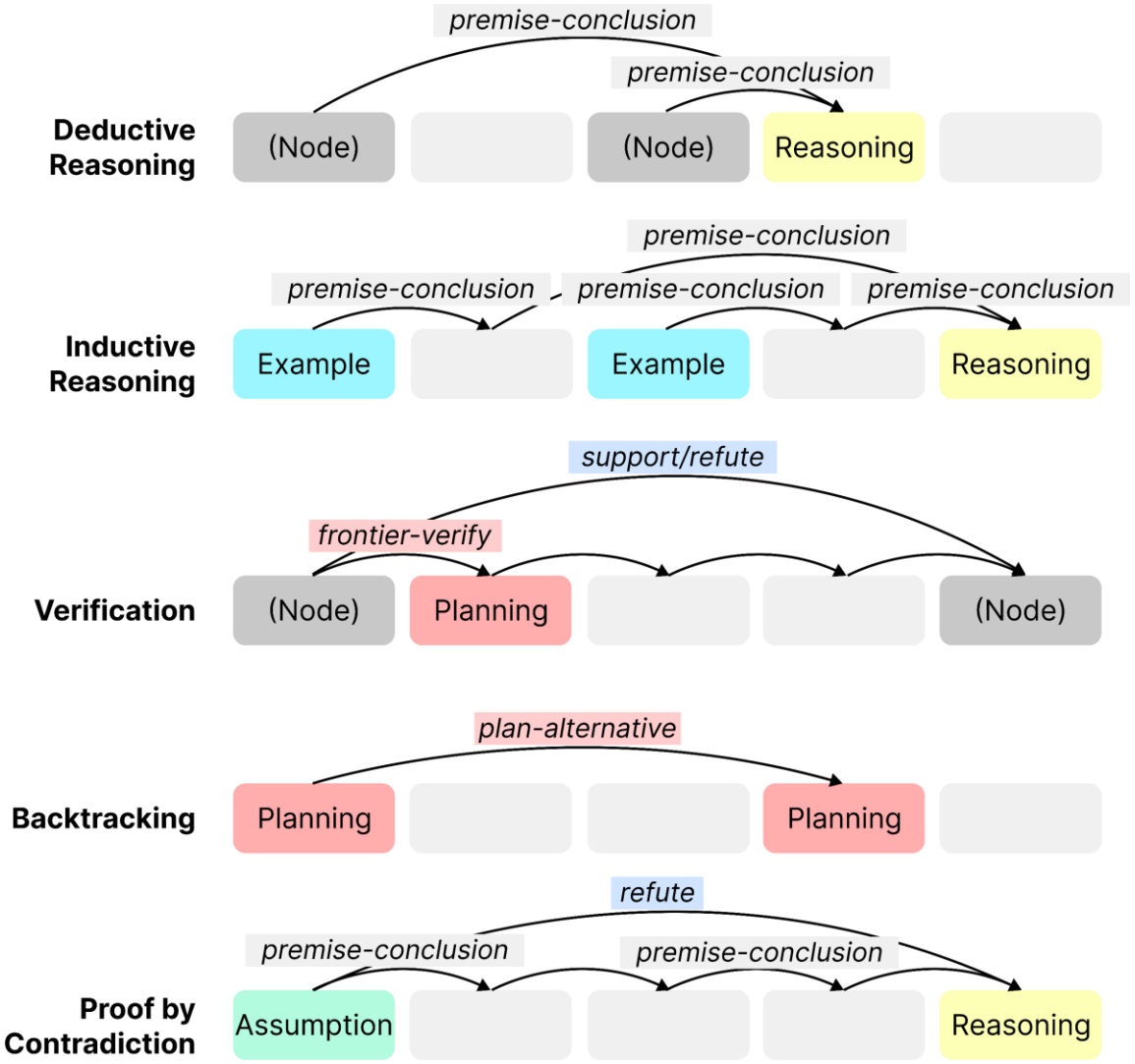
Example ReasoningFlow graph



ReasoningFlow labels

Reasoning behaviors emerge as graph patterns

- Previous works identify specific behaviors (Deduction, Induction, Self-verification, ...)
- With ReasoningFlow, we can formally define these behaviors by graph patterns



Future works

1. Scaling up ReasoningFlow: More data, Fast automatic parser

- Currently, 30 traces with manual annotations are released
- Applying LLM-based annotation?
  - LLMs are powerful in node annotation pipeline (segmentation + classification)
  - However, LLMs seem to suffer on edge annotation (detection + classification)

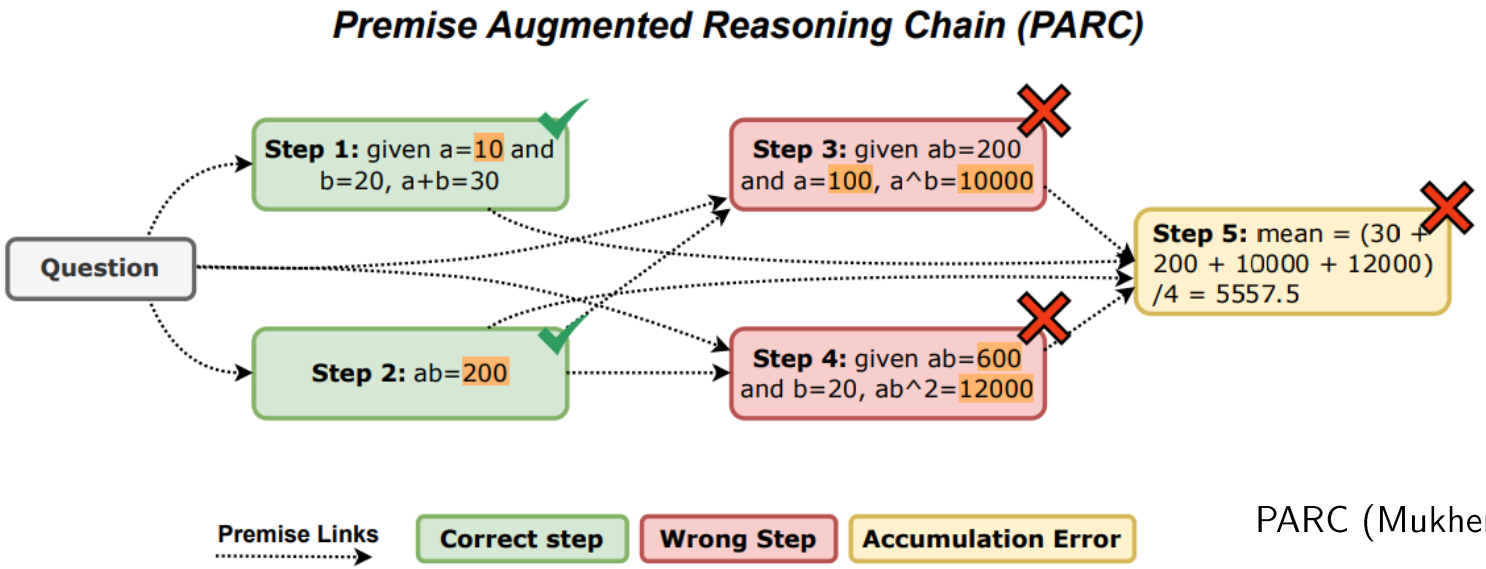
Model	Node seg.	Node class.	Edge det.	Edge class.
GPT-5-mini	0.72	0.81	0.31	0.37
Gemini-2.5-Flash	0.79	0.84	0.43	0.30

F1 scores for 4 stages of automatic ReasoningFlow annotation.

Future works

2. Applications: Verification with partial context

- Verifying long reasoning traces is hard
  - Existing methods only target **first error**, ignoring error propagation
  - Cannot directly apply PRMs/LLM-as-a-judge due to non-linearity/length
- Use partial context for step-wise verification
  - Can track **error propagation**
  - Can apply existing approaches with pruned context



PARC (Mukherjee et al., 2025)

Future works

2. Applications: Tracking effects of RAG documents

- Not all RAG contexts are used for retrieval-augmented reasoning
- Can we track which documents were used for reasoning and which were not?
  - Further training signals for retriever-for-reasoning
  - Better compensation system for authors