Thesis Proposal: Interpretable Reasoning Enhancement in Large Language Models through Puzzle and Ontological Task Analysis

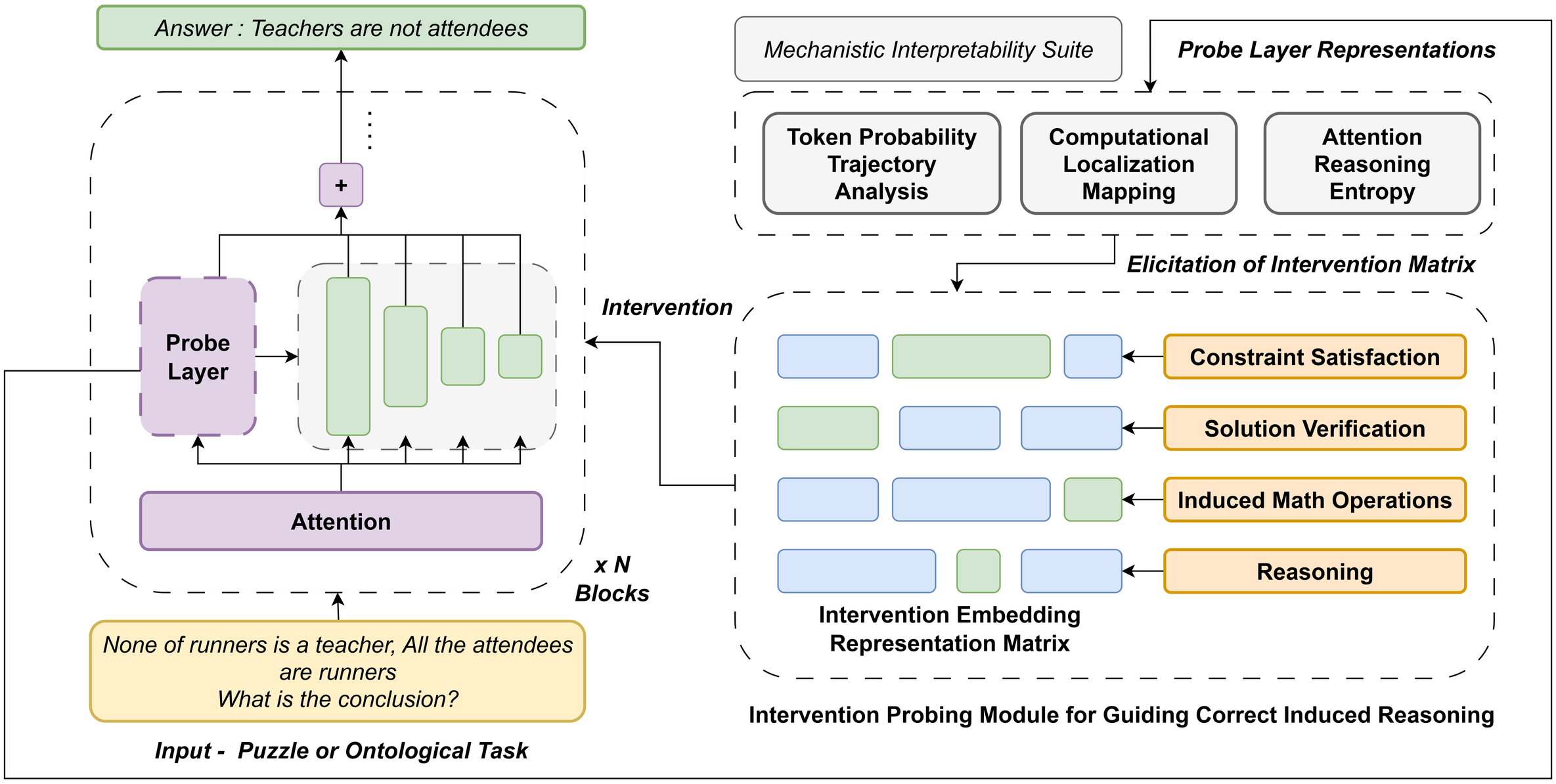
Mihir Panchal

IJCNLP-AACL 2025

Research Questions

How can systematic error patterns in domain specific reasoning be detected through layer wise probing and mitigated through targeted interventions?

How can probing frameworks and middle layer analyses reveal and enhance the computational mechanisms underlying inference?



Aims and Timeline Deliverables

Year 1 — Foundation

- Literature review on LLM reasoning & interpretability
- Develop pilot datasets
- Build initial probing architectures for puzzle ontological tasks

Year 2 — Scaling & Validation

- Complete full annotated datasets
- Refine & validate probing classifiers
- Conduct middle layer analysis across different model architectures

Year 3 — Pattern Discovery

- Identify reasoning representation patterns across domains
- Develop cross task pattern discovery methods
- Design probe guided intervention strategies

Year 4 — Implementation & Completion

- Validate intervention systems with real time enhancement
- Large scale experiments & human evaluation
- Thesis writing & defense preparation

Pilot datasets, baseline probing architectures, 1-2 publications

Public dataset release, validated classifiers, 1-2 conference papers

Unified analysis framework, intervention prototypes, 1-2 major publications

Complete intervention system, thesis, final publications, public tools

Scan here to connect or collaborate

Scan here to read the thesis proposal