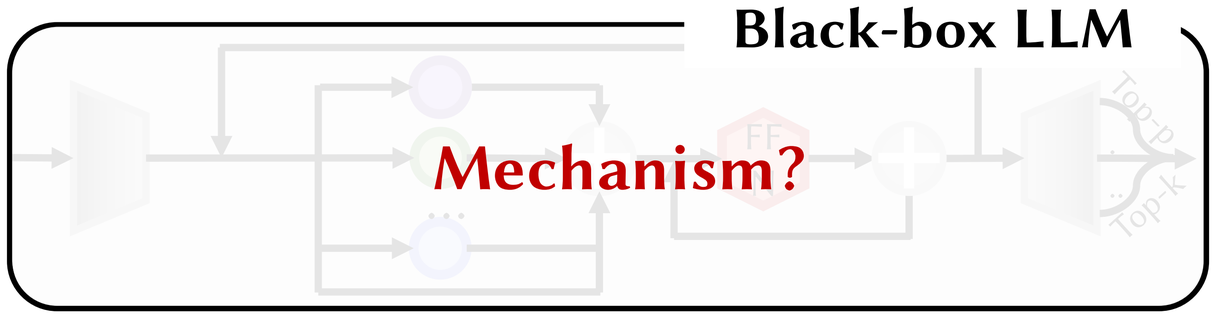
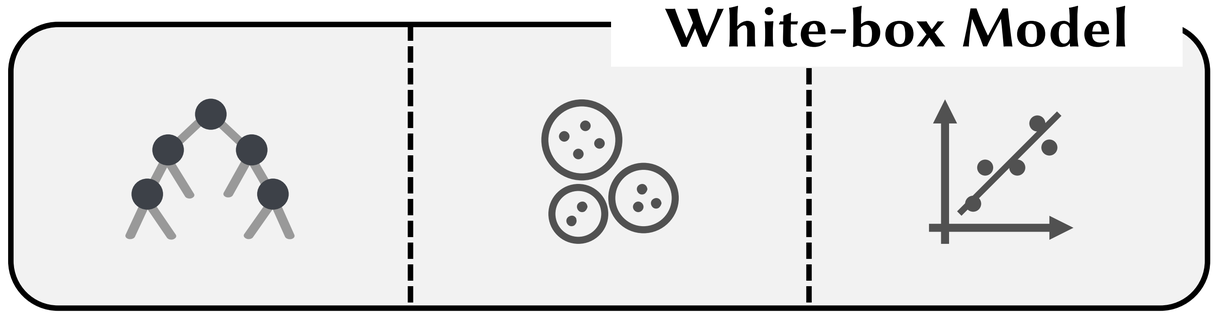
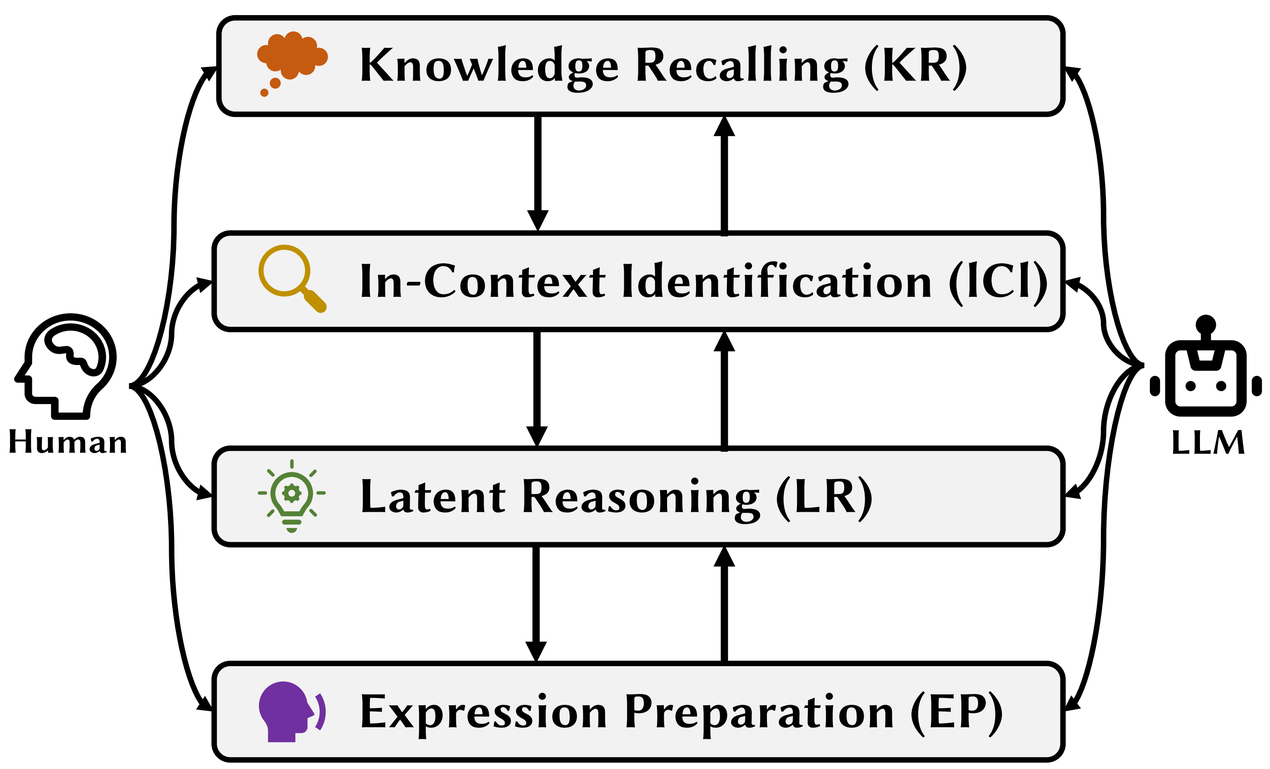
Attention Heads of Large Language Models: A Survey

Introduction & Background

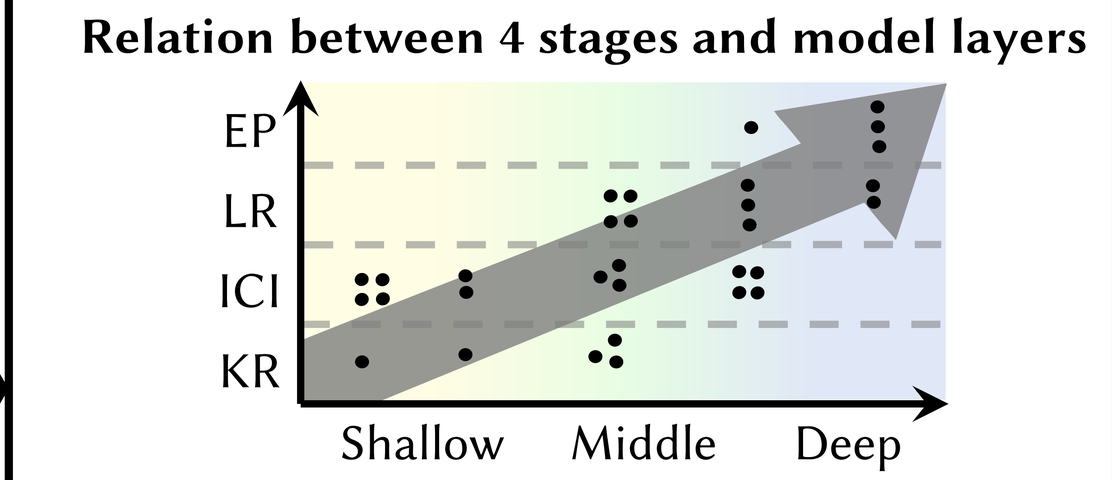




Special Attention Heads



Collaborative Mechanism



Discussion

Limitations

- Generalizability
- Multi-collaboration
- Theoretical support

Future Directions

- Complex Task
- Unified Framework
- Align with Human

Evaluation

- Mechanism Exploration
- Common Evaluation

Add. Topics

- FFN Interpretability
- Machine Psychology

Experimental Method