Attention Heads of Large Language Models: A Survey

Introduction & Background (Sec.1-2)

Black-box LLM

Mechanism?

White-box Model

Special Attention Heads (Sec.3.1–3.5)

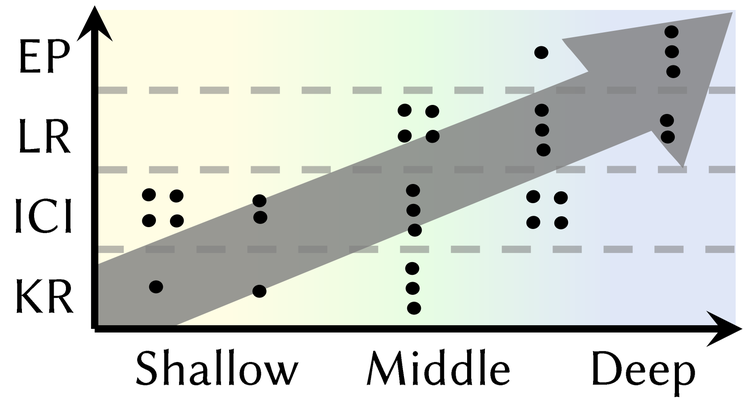
- Knowledge Recalling (KR)
- In-Context Identification (ICI)
- Latent Reasoning (LR)
- Expression Preparation (EP)

Human

LLM

Collaborative Mechanism (Sec.3.6)

Relation between 4 stages and model layers



Application Scene

- MCQA
- Parity Problem
- IOI
- ...

Experimental Method (Sec.4)

- Modeling-Free
  - Modification-Based
  - Replacement-Based
- Modeling-Required
  - Training-Free
  - Training-Required

Evaluation (Sec.5)

- Mechanism Exploration
- Common Evaluation

Add. Topics (Sec.6)

- FFN Interpretability
- Machine Psychology

Conclusion (Sec.7)

Limitations

- Universality
- Theoretical support

Future Directions

- Complex Task
- Unified Framework
- Align with Human