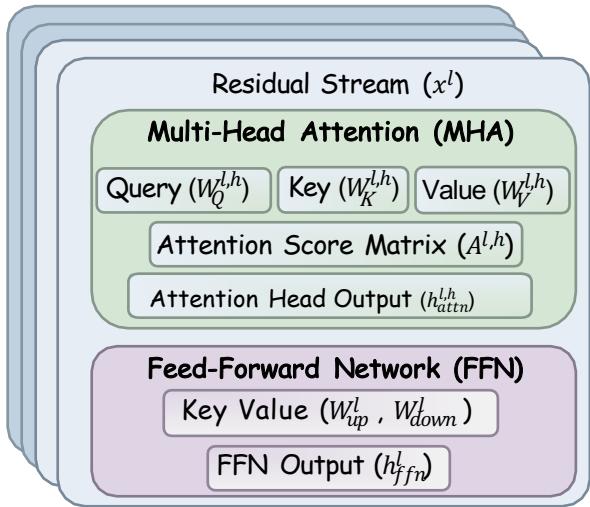


Interpretable Objects (Decoder-only Transformer LLM)

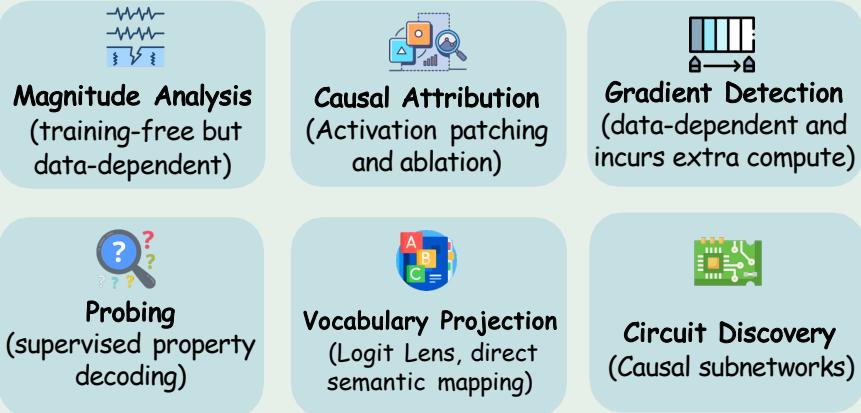


Other Specific Objects

- Neurons/Activations (s^l, k_j^l, v_j^l)
- Sparse Autoencoder (SAE) Features (f_j, a_j)
- Circuits/Subnetworks (O, C)

Methods for Localizing and Steering

Localizing Methods (Identifying & Localizing Objects)



Steering Methods (Controlling & Improving Model)



Applications (Model Improvement)

Improve Alignment

- Safety and Reliability
- Fairness and Bias
- Persona and Role

Improve Capability

- Multilingual & Crosslingual
- Knowledge Management
- Logic and Reasoning

Improve Efficiency

- Efficient Training
- Efficient Inference

► Actionable Mechanistic Interpretability: From Localizing and Steering to Model Improvement