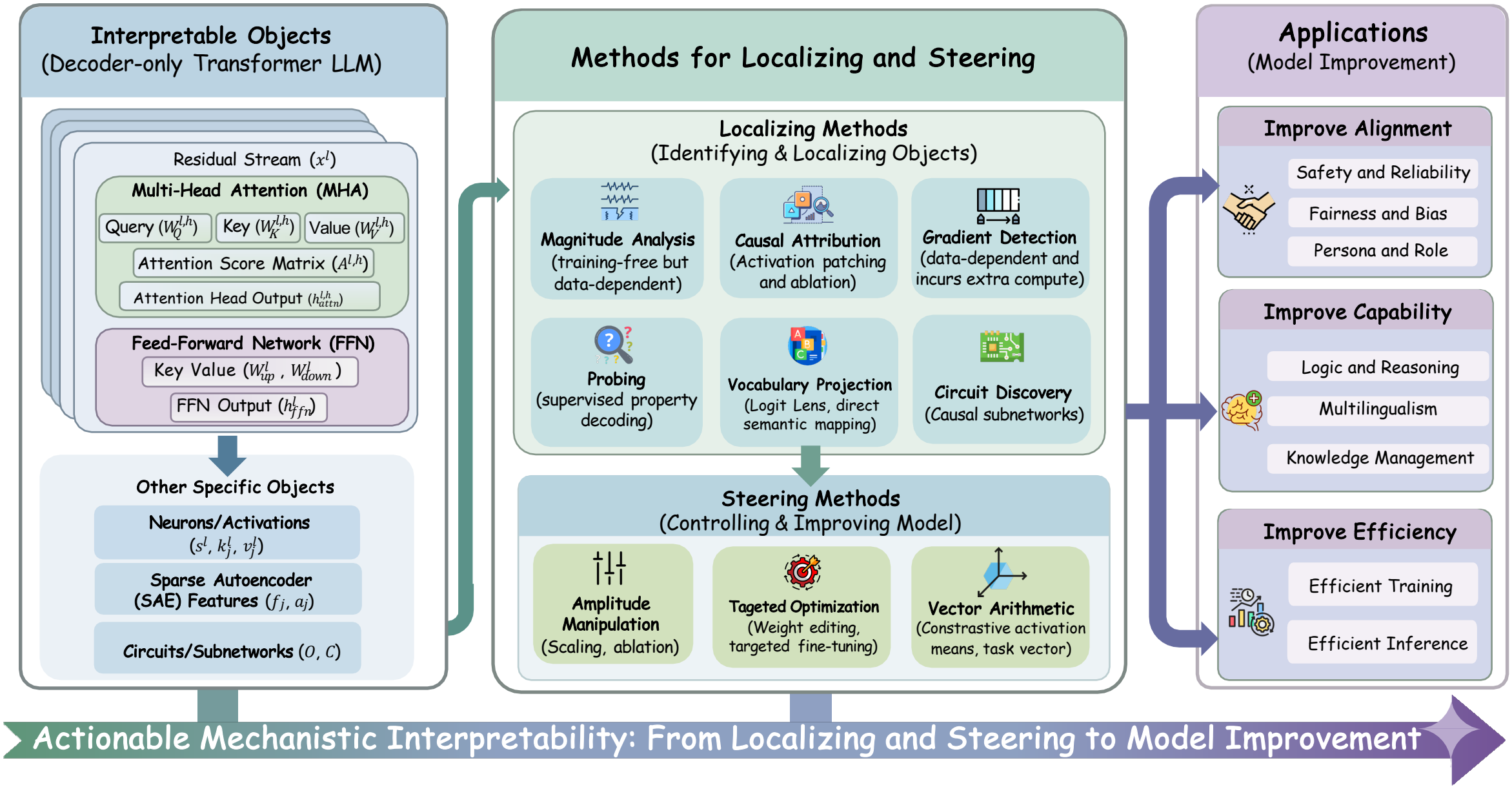
Interpretable Objects
(Decoder-only Transformer LLM)
Residual Stream (x^l)
Multi-Head Attention (MHA)
Query ($W_Q^{l,h}$)
Key ($W_K^{l,h}$)
Value ($W_V^{l,h}$)
Attention Score Matrix ($A^{l,h}$)
Attention Head Output ($h_{attn}^{l,h}$)
Feed-Forward Network (FFN)
Key Value (W_{up}^l , W_{down}^l)
FFN Output (h_{ffn}^l)
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)
Magnitude Analysis (training-free but data-dependent)
Causal Attribution (Activation patching and ablation)
Gradient Detection (data-dependent and incurs extra compute)
Probing (supervised property decoding)
Vocabulary Projection (Logit Lens, direct semantic mapping)
Circuit Discovery (Causal subnetworks)
Steering Methods
(Controlling & Improving Model)
Amplitude Manipulation (Scaling, ablation)
Tageted Optimization (Weight editing, targeted fine-tuning)
Vector Arithmetic (Constrastive activation means, task vector)
Applications
(Model Improvement)
Improve Alignment
Safety and Reliability
Fairness and Bias
Persona and Role
Improve Capability
Logic and Reasoning
Multilingualism
Knowledge Management
Improve Efficiency
Efficient Training
Efficient Inference
Actionable Mechanistic Interpretability: From Localizing and Steering to Model Improvement