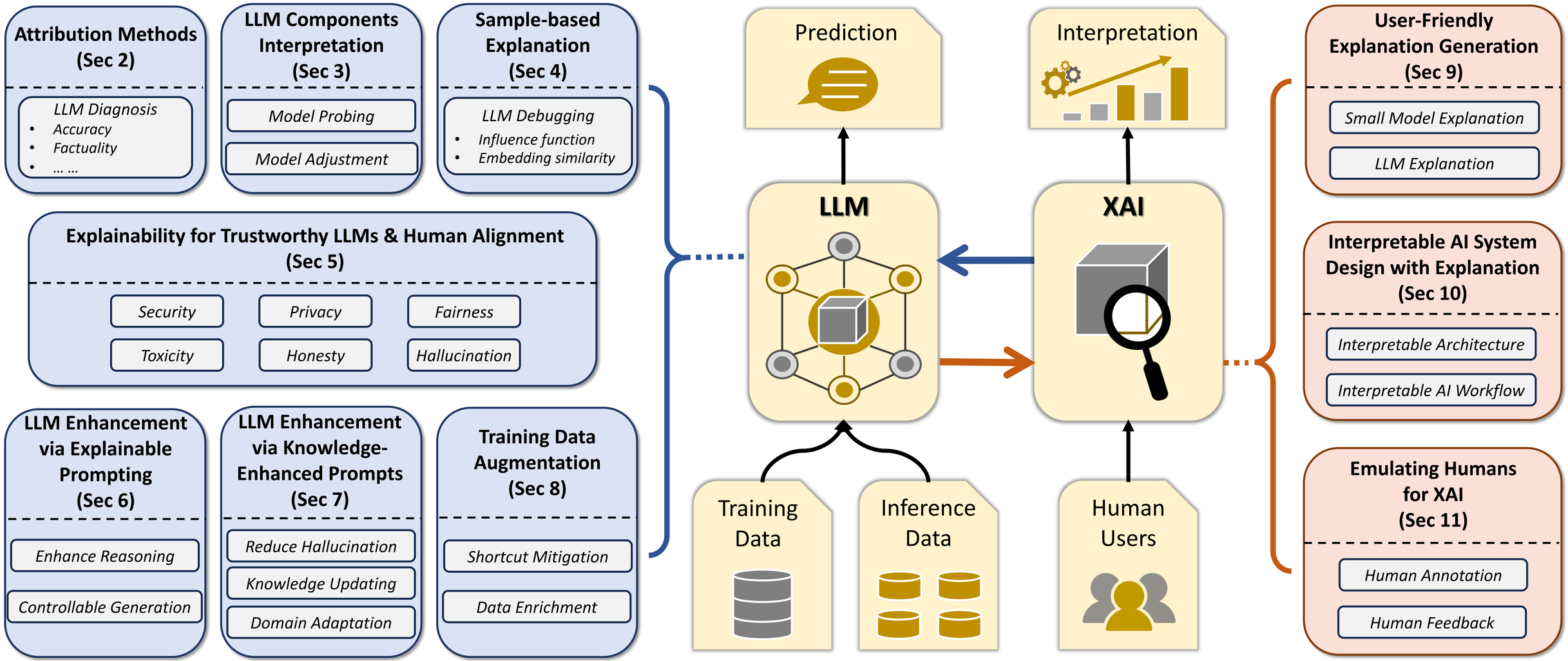
Attribution Methods (Sec 2)
LLM Diagnosis
Accuracy
Factuality
… …
LLM Components Interpretation (Sec 3)
Model Probing
Model Adjustment
Sample-based Explanation (Sec 4)
LLM Debugging
Influence function
Embedding similarity
Explainability for Trustworthy LLMs & Human Alignment (Sec 5)
Security
Privacy
Fairness
Toxicity
Honesty
Hallucination
LLM Enhancement via Explainable Prompting (Sec 6)
Enhance Reasoning
Controllable Generation
LLM Enhancement via Knowledge-Enhanced Prompts (Sec 7)
Reduce Hallucination
Knowledge Updating
Domain Adaptation
Training Data Augmentation (Sec 8)
Shortcut Mitigation
Data Enrichment
Prediction
Interpretation
LLM
XAI
Training Data
Inference Data
Human Users
User-Friendly Explanation Generation (Sec 9)
Small Model Explanation
LLM Explanation
Interpretable AI System Design with Explanation (Sec 10)
Interpretable Architecture
Interpretable AI Workflow
Emulating Humans for XAI (Sec 11)
Human Annotation
Human Feedback