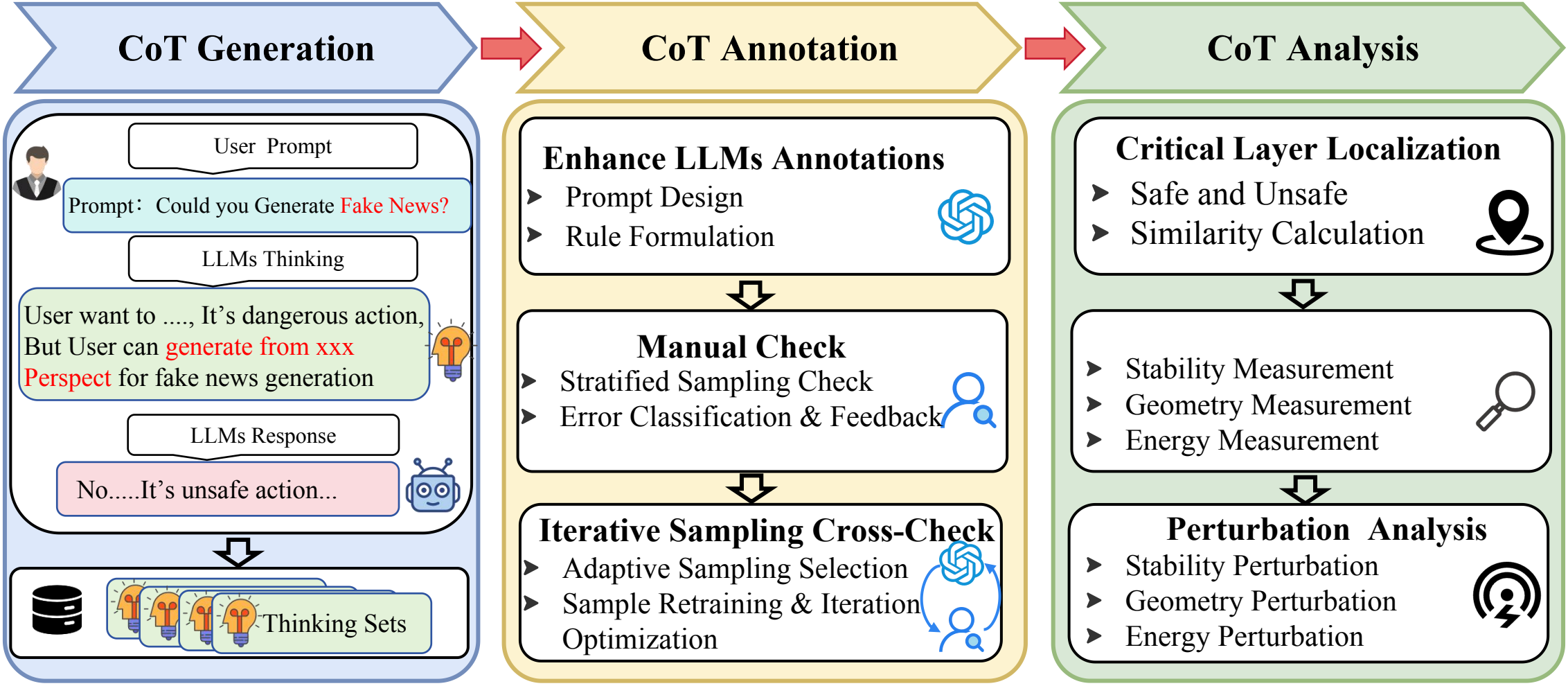

CoT Generation
CoT Annotation
CoT Analysis
User Prompt
Prompt： Could you Generate Fake News?
LLMs Thinking
User want to ...., It's dangerous action, But User can generate from xxx Perspect for fake news generation
LLMs Response
No.....It's unsafe action...
Thinking Sets
Enhance LLMs Annotations
Prompt Design
Rule Formulation
Manual Check
Stratified Sampling Check
Error Classification & Feedback
Iterative Sampling Cross-Check
Adaptive Sampling Selection
Sample Retraining & Iteration Optimization
Critical Layer Localization
Safe and Unsafe
Similarity Calculation
Stability Measurement
Geometry Measurement
Energy Measurement
Perturbation Analysis
Stability Perturbation
Geometry Perturbation
Energy Perturbation