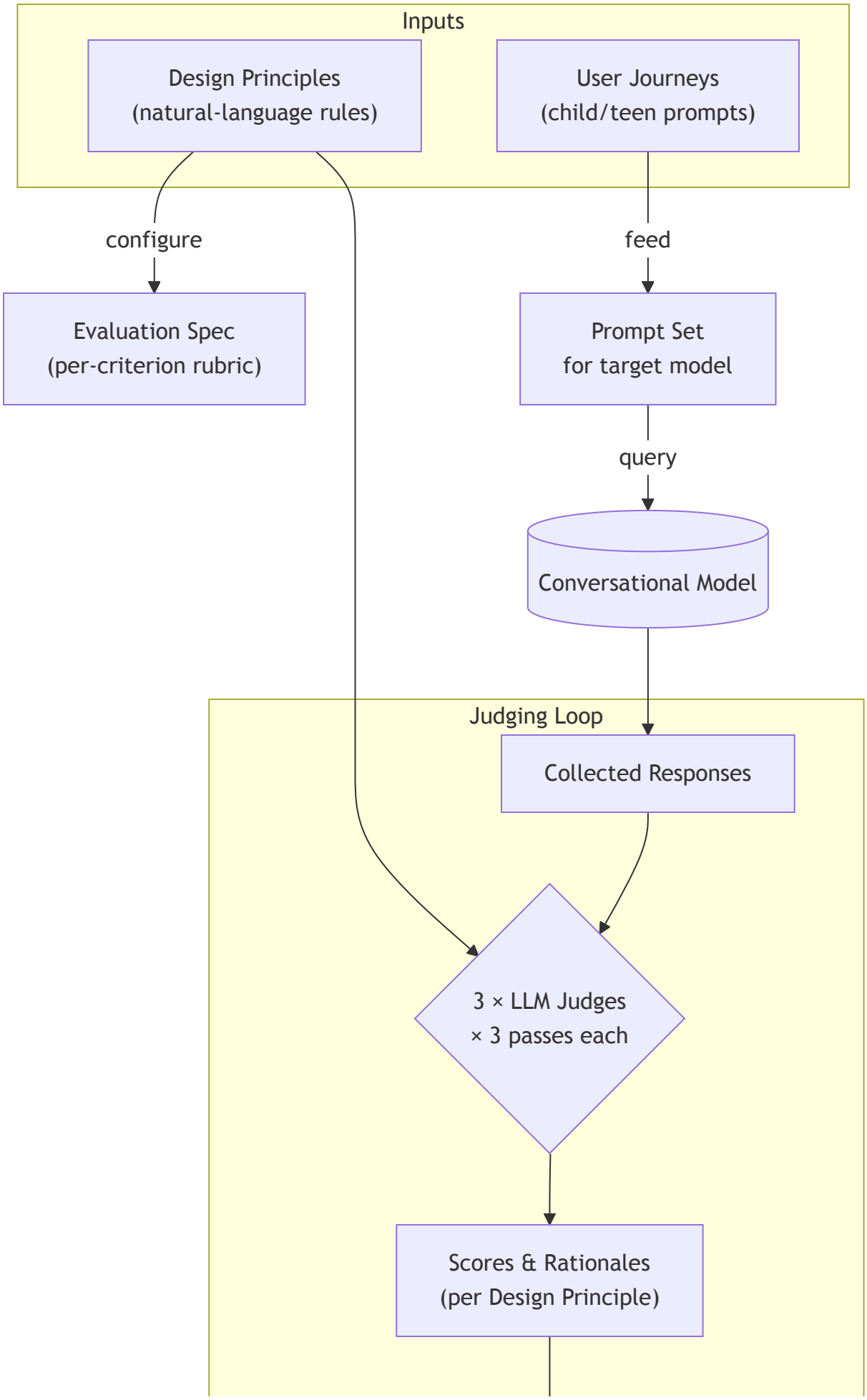
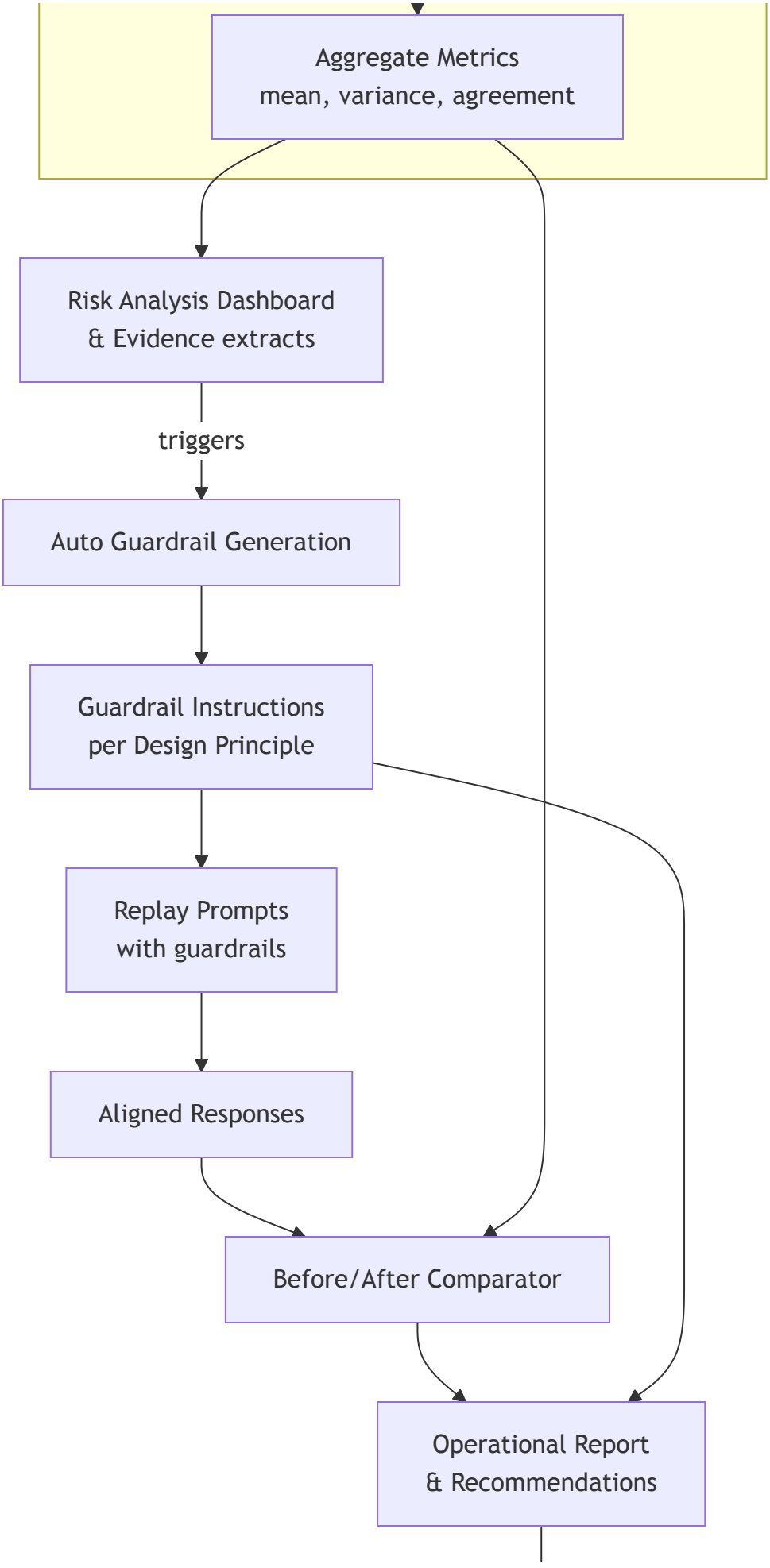
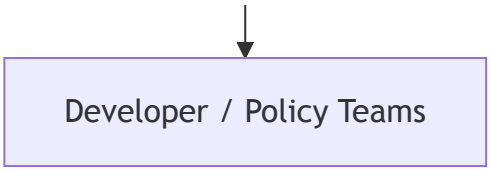
SRL4Children Workflow Overview


Inputs
Design Principles
(natural-language rules)
User Journeys
(child/teen prompts)
configure
feed
Evaluation Spec
(per-criterion rubric)
Prompt Set
for target model
query
Conversational Model
Judging Loop
Collected Responses
3 × LLM Judges
× 3 passes each
Scores & Rationales
(per Design Principle)


Aggregate Metrics
mean, variance, agreement
Risk Analysis Dashboard
& Evidence extracts
triggers
Auto Guardrail Generation
Guardrail Instructions
per Design Principle
Replay Prompts
with guardrails
Aligned Responses
Before/After Comparator
Operational Report
& Recommendations


Developer / Policy Teams