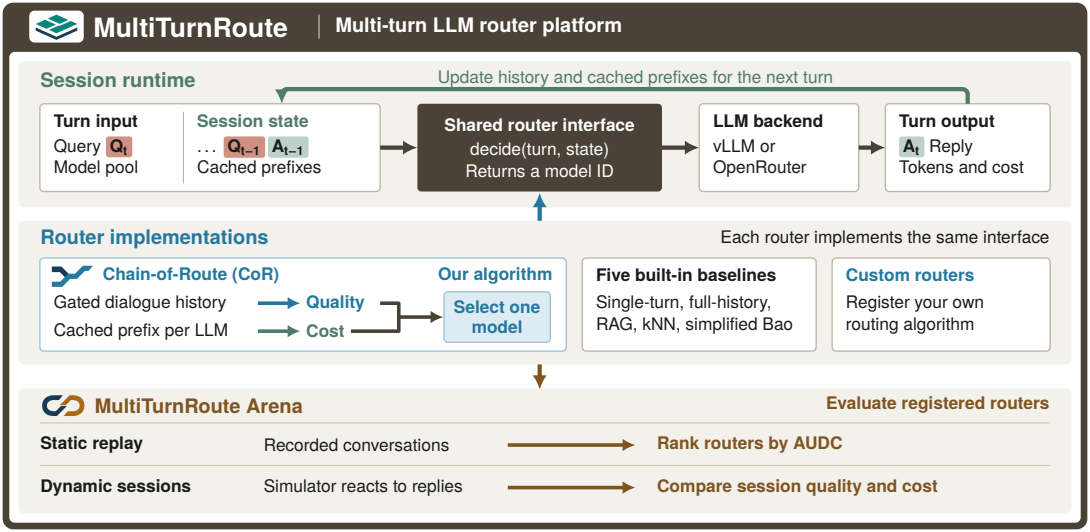

MultiTurnRoute
Multi-turn LLM router platform
Session runtime
Update history and cached prefixes for the next turn
Turn input
Query Q_t
Model pool
Session state
... Q_{t-1} A_{t-1}
Cached prefixes
Shared router interface
decide(turn, state)
Returns a model ID
LLM backend
vLLM or
OpenRouter
Turn output
A_t Reply
Tokens and cost
Router implementations
Each router implements the same interface
Chain-of-Route (CoR)
Our algorithm
Gated dialogue history
Cached prefix per LLM
Quality
Cost
Select one model
Five built-in baselines
Single-turn, full-history, RAG, kNN, simplified Bao
Custom routers
Register your own routing algorithm
MultiTurnRoute Arena
Evaluate registered routers
Static replay
Recorded conversations
Rank routers by AUDC
Dynamic sessions
Simulator reacts to replies
Compare session quality and cost