Branching Memory for AI Conversations

A Six-Layer Architecture

1. Graph instead of a flat log

Conversation history is an immutable graph.
Each exchange: node_id, parent_id,
branch_id

2. Context assembly as a graph path

Not "what to truncate", but "which path
through the graph" to pick for the query

3. Incremental replication

Client-server sync without resending
the full history (offline-first)

4. Visualization and navigation

User can see and navigate
branches of the conversation directly

5. Model-agnostic compute layer

Graph logic is independent
of the specific LLM used

6. Integrity-check layer (pollinator)

Independent, continuously running check
of graph structure — outside
the orchestrator and the model