Traditional RAG

User → Query → **Retrieve** → **Augment** (Prompt) → **Generate** (OpenAI, Gemini, AI)
Response ← (back to User)

RAG in a System 2 view

User → Query, Budgets (time, cost, quality ,...) → **Plan** (Reasoning and decomposition) →

Example Plans

Discover (Schema discovery) → **Retrieve** (NL2SQL -> SQL query)

or

Discover (Data discovery) → **Retrieve** (Vector search + graph search)

or

Discover (Data discovery) → **Retrieve** (Vector + graph + web searches)

, and more alternatives...

→ **Augment** (Prompt) → **Generate** (OpenAI, Gemini, AI) → **Scrutinize** (Verification) → Response → (back to User)

Replan as needed (Scrutinize → Plan)