

# A Mental Model

Inconsistent MCP behavior is almost never a model problem. It's a design problem — and it starts with the wrong mental model.

APIs are designed for  
**EXECUTION**

vs

MCP Servers must support  
**REASONING BEFORE EXECUTION**

## THE BROKEN ASSUMPTION

### API design patterns that fail LLMs

- Mirror the API surface
- Expose flexible operations
- Rely on implicit behavior
- Assume the consumer will adapt

## BEFORE CALLING ANY TOOL, THE LLM MUST...

- 01 Decide whether a tool should be used
- 02 Choose which tool applies
- 03 Determine how to populate inputs
- 04 Predict what will happen
- 05 Interpret the result
- 06 Decide what to do next

## WHAT AN LLM ACTUALLY SEES

### Its entire world:

- Tool names
- Descriptions
- Parameters
- Outcomes

If something isn't made explicit here, it effectively doesn't exist.

## 4 QUESTIONS EVERY TOOL MUST ANSWER

- Can the LLM decide when to use this?
- Can it distinguish this from other options?
- Can it predict what will happen?
- Can it understand the result well enough to continue?

If your MCP Server feels unpredictable, ask: was this designed for **execution** — or for **reasoning**?