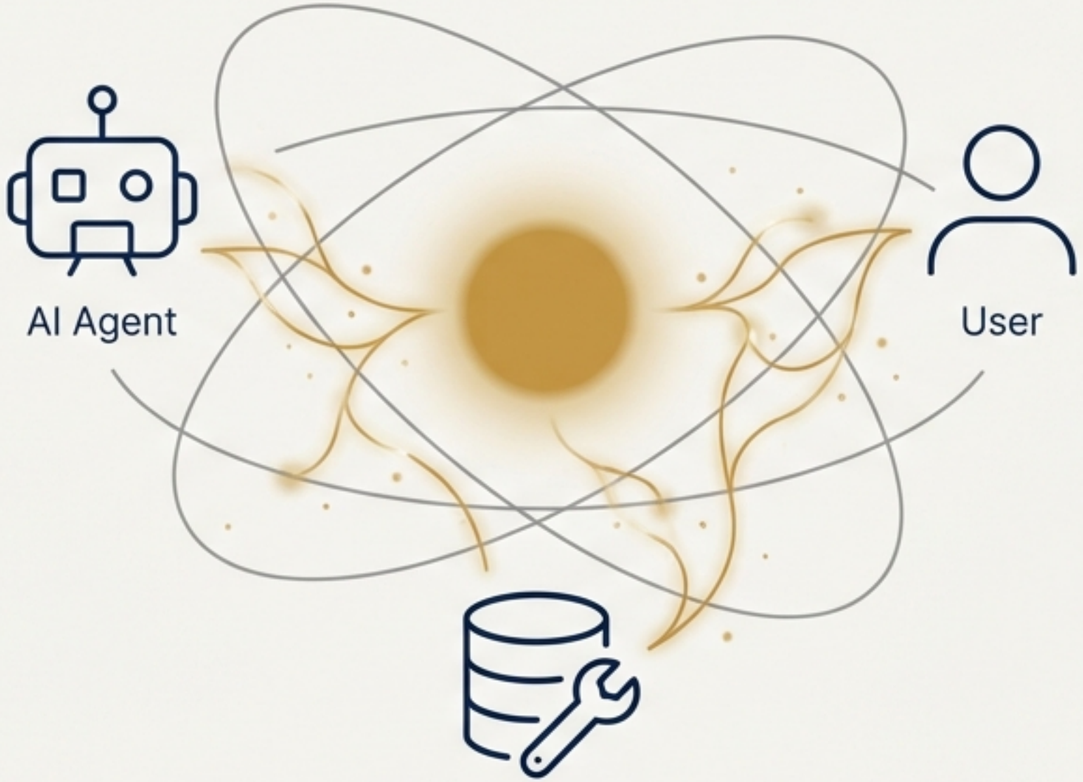
Beyond Tool Use: Benchmarking Agents for Real-World Collaboration

The evolution from τ-bench to τ2-bench in the quest for realistic agent evaluation.



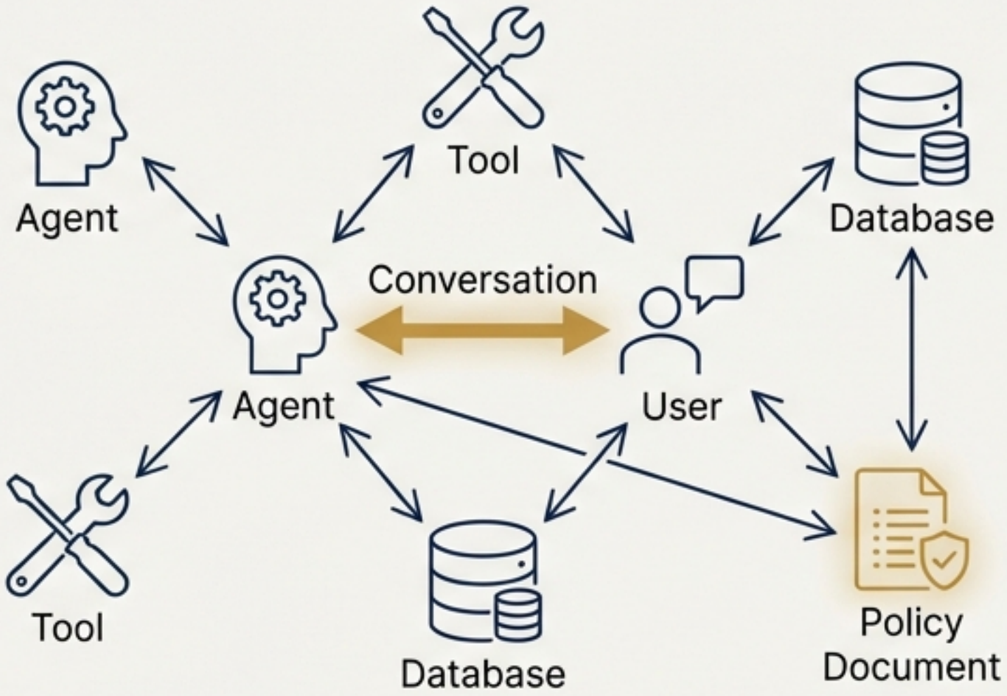
Today's Agent Benchmarks Are Missing What Matters: Users and Rules

Most benchmarks feature simplified instruction-following setups, ignoring three critical desiderata for deploying agents in the wild:

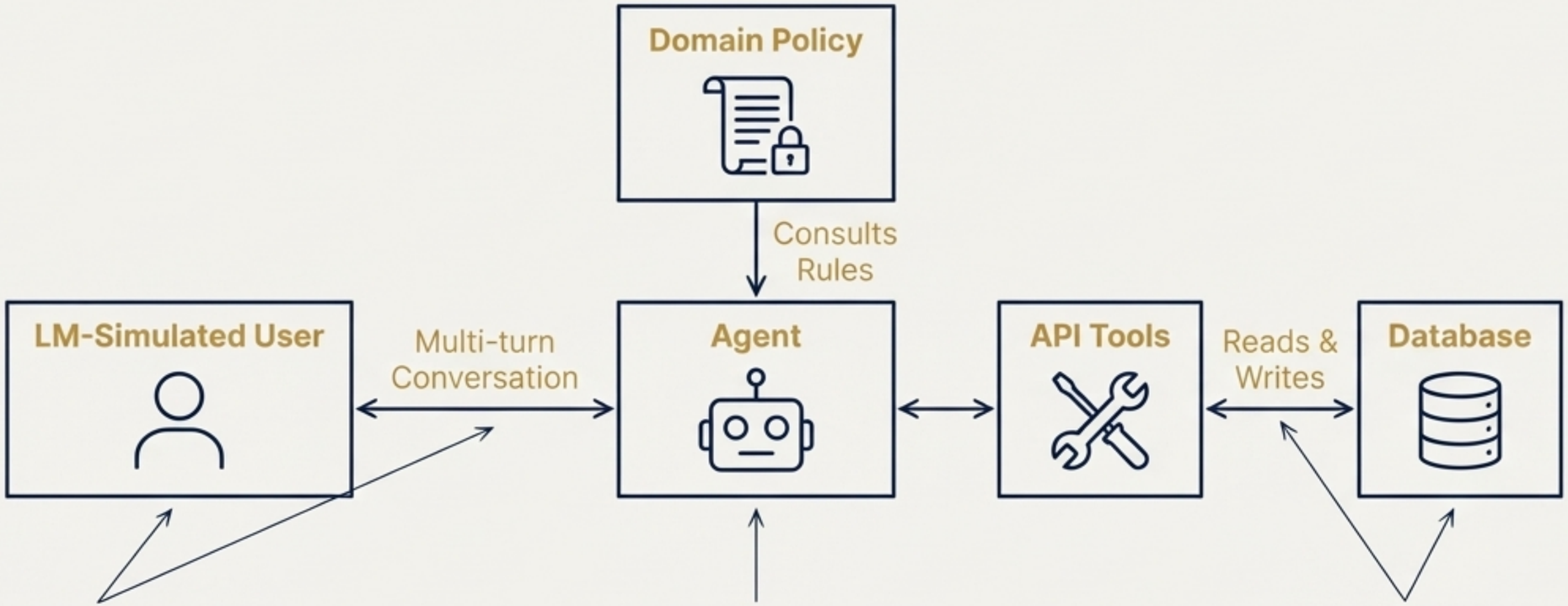
1. Human Interaction: Agents must incrementally gather information and resolve intents through dynamic, multi-turn conversations.
2. Rule Adherence: Agents must accurately adhere to complex policies and rules specific to a task or domain.
3. Consistency: Agents must maintain reliability at scale, across millions of stochastic interactions.

Typical Benchmark

Real World



τ-bench Simulates Real-World Dynamics: Users, Tools, and Policies



LM-Simulated User

An LLM simulates a human user, creating stochastic, multi-turn conversations to test the agent's interactive capabilities.

Domain Policies

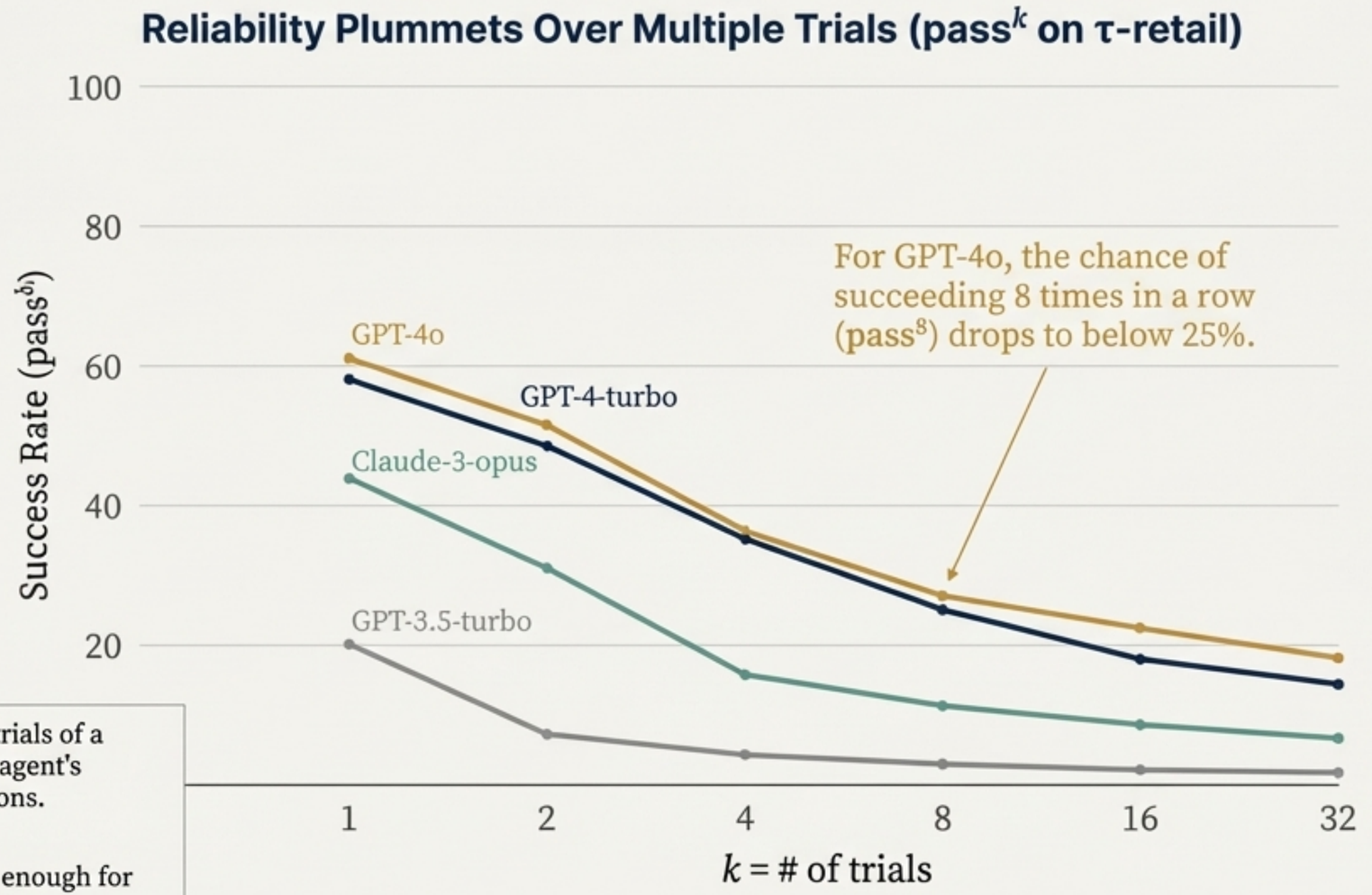
Agents are provided with policy documents (e.g., airline cancellation rules) that they must consult and adhere to.

Stateful Evaluation

Success is measured by comparing the final database state to the ground truth, allowing varied conversational paths to the same correct outcome.

The First Revelation: State-of-the-Art Models Are Strikingly Inconsistent

On τ-bench, even top models like GPT-4o achieve low task success rates (~61% on τ-retail and ~35% on τ-airline).

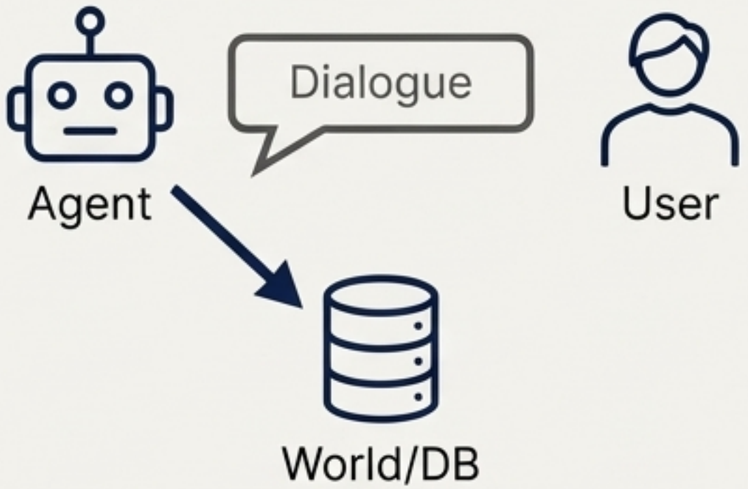


pass^k is the chance that all k i.i.d. trials of a task are successful. It measures the agent's robustness to conversational variations.

Takeaway:
High average success (pass^1) is not enough for real-world deployment; agents must be consistent.

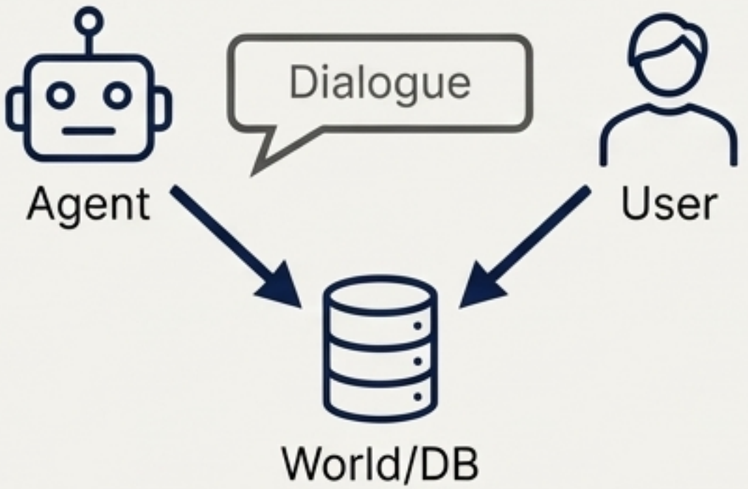
But τ-bench Has a Limitation: What If the User Needs to Act?

Single-Control: τ-bench



The agent is the sole actor. The user is a passive information source (e.g., booking a flight).

Dual-Control: Real World



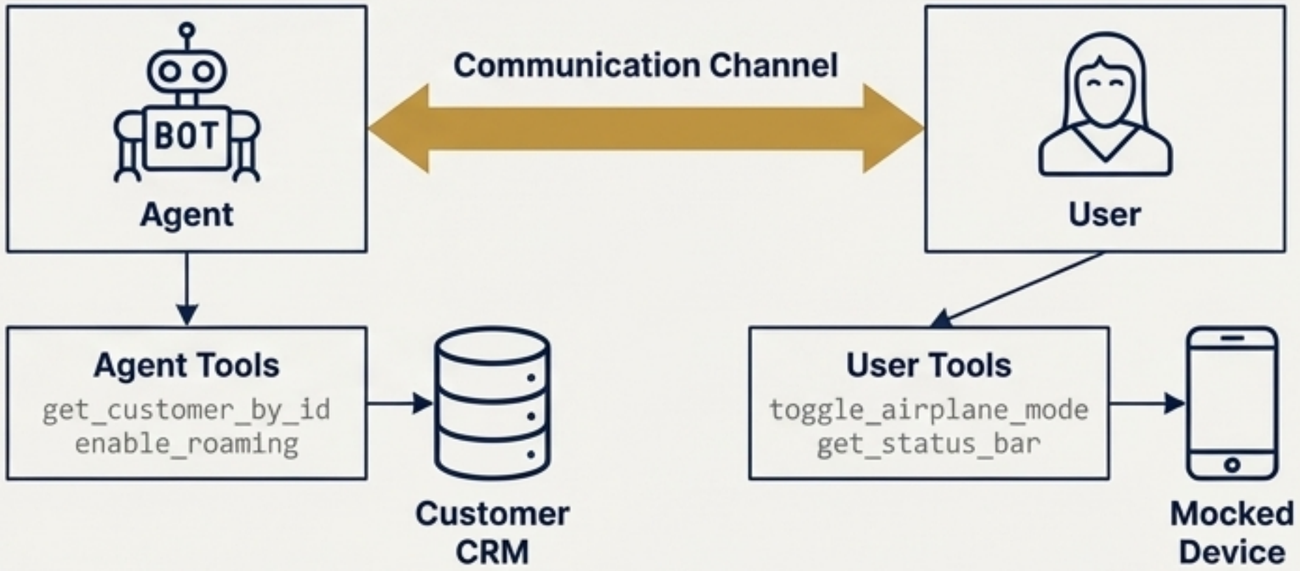
Both agent and user must act to solve the problem, requiring coordination (e.g., technical support).

How can we evaluate an agent's ability to not just act, but to guide a user to act correctly?

τ2-bench introduces a Dual-Control Environment for Collaborative Tasks

Core Innovation

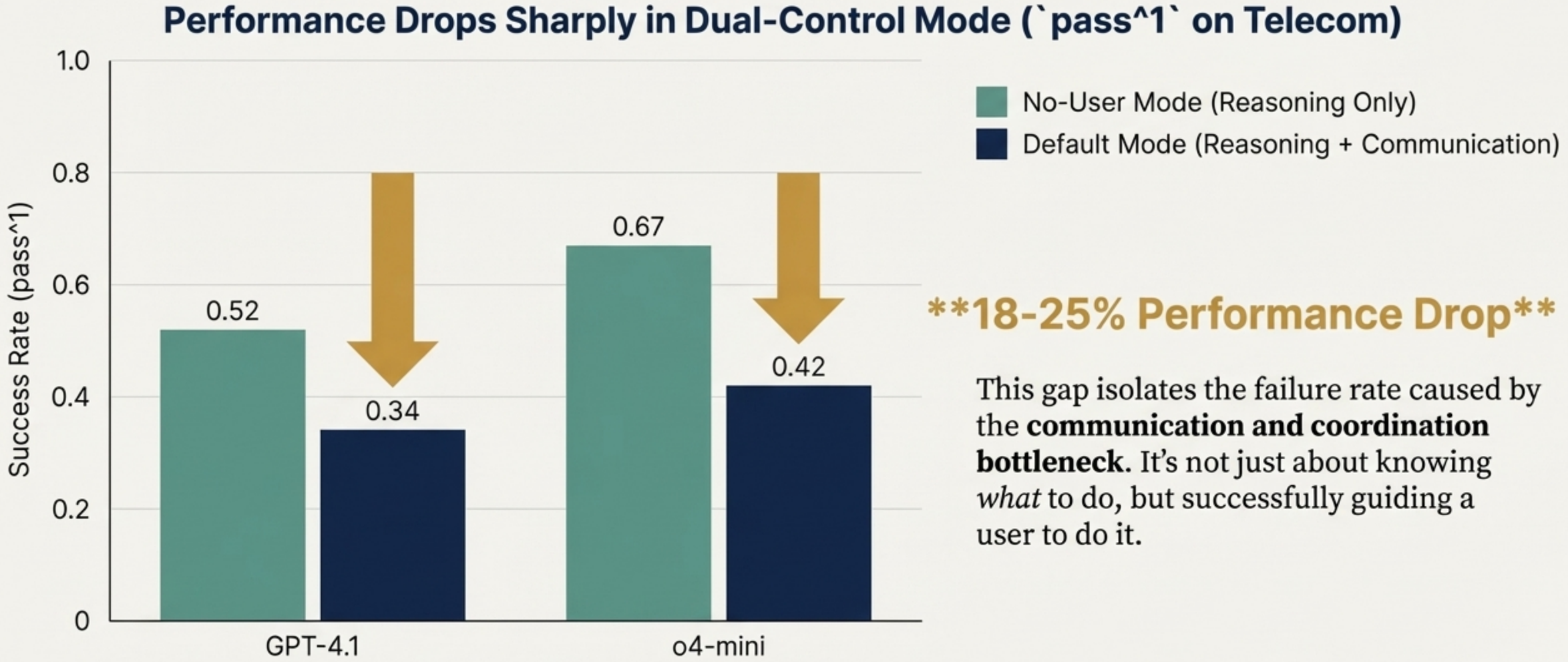
Modeled as a **Decentralized Partially Observable Markov Decision Process (Dec-POMDP)**, where both agent and user have distinct tools to act on a shared, dynamic environment.



New Domain: Telecom Troubleshooting

- **The Scenario:** An agent helps a user diagnose and fix a phone issue (e.g., no mobile data).
- **Agent Tools:** Access and modify customer CRM data (e.g., `get_customer_by_id`, `enable_roaming`).
- **User Tools:** Interact with a mocked phone device (e.g., `toggle_airplane_mode`, `get_status_bar`).
- **The Challenge:** The agent must successfully diagnose the issue and guide the user to perform the correct sequence of actions on their end.

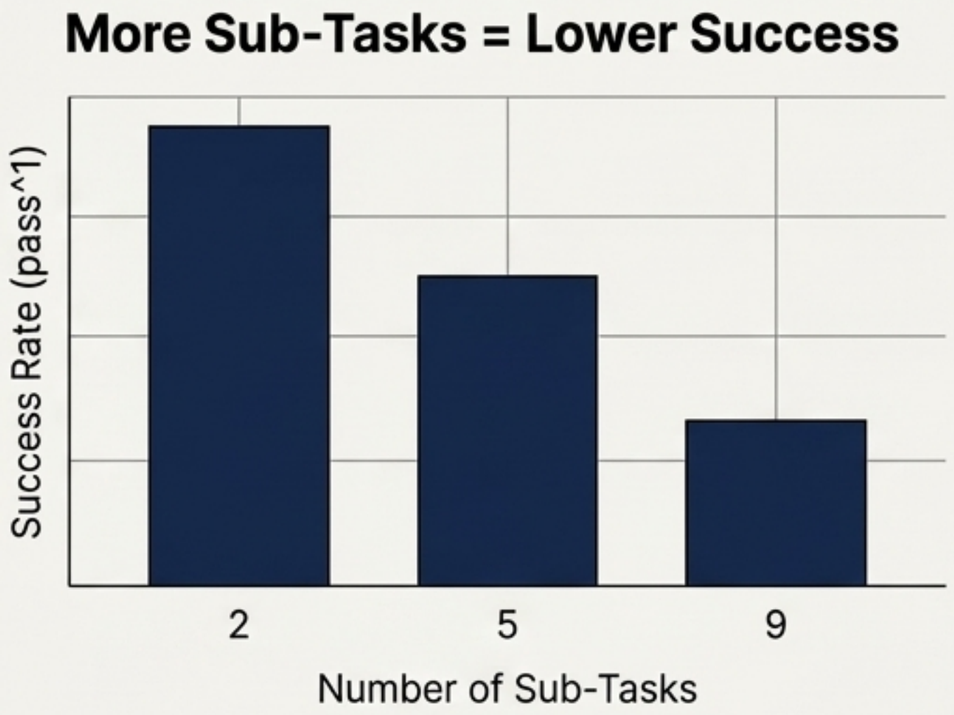
The Second Revelation: Guiding a User Is Harder Than Acting Alone

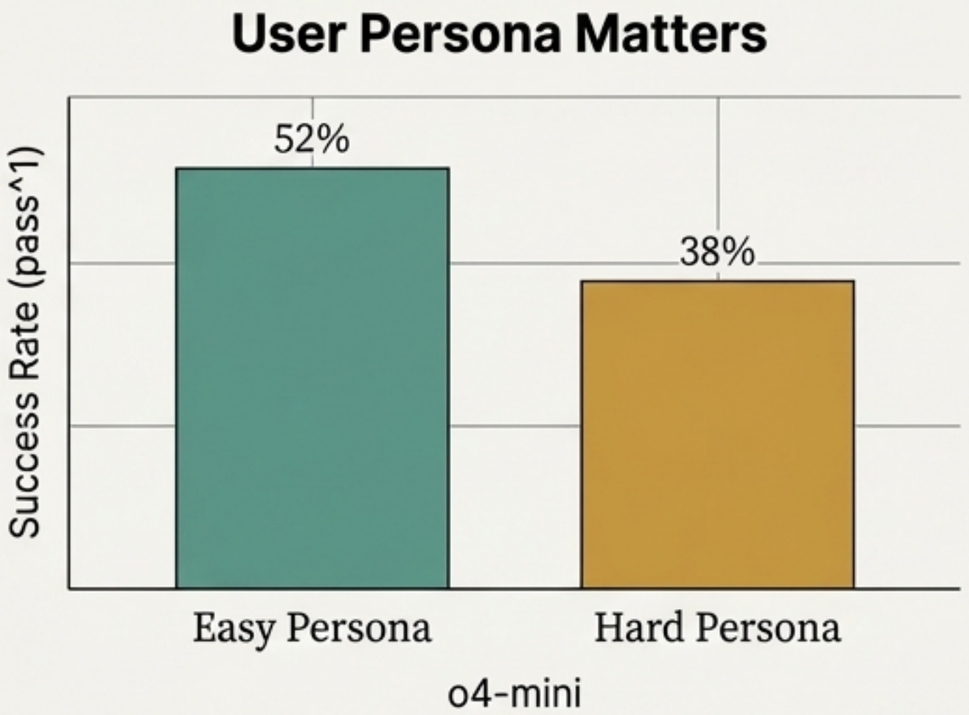


****18-25% Performance Drop****

This gap isolates the failure rate caused by the **communication and coordination bottleneck.** It's not just about knowing *what* to do, but successfully guiding a user to do it.

Diagnosis: Task Complexity and User Persona Amplify the Challenge



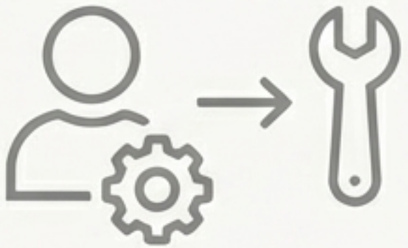


τ2-bench's compositional tasks and user personas enable fine-grained diagnosis, moving beyond a single success score to understand *why* agents fail.

The Quest for Realism: An Evolutionary View

Increasing Realism

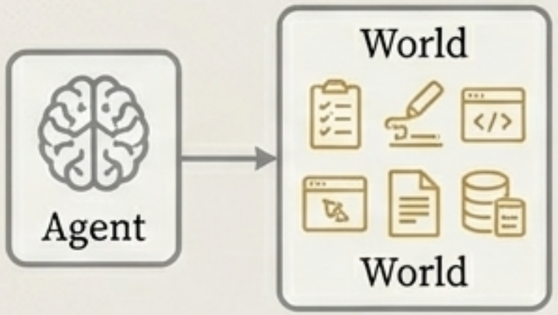
Traditional Benchmarks



Instruction Following

No user interaction, no domain rules, no measure of consistency.

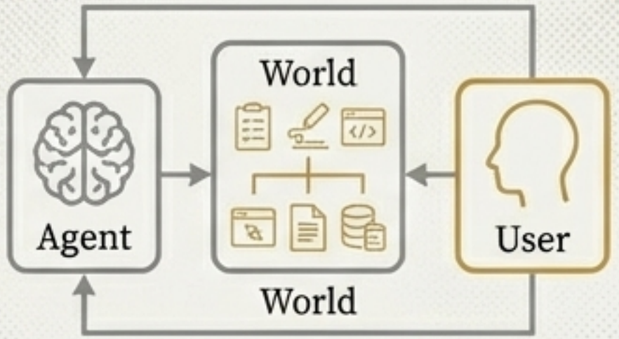
τ-bench



Single-Control

Exposed the critical **inconsistency** of SOTA models via `pass^k.

τ2-bench



Dual-Control

Isolated the **communication bottleneck** as a major failure point.

The Path Forward: Building Agents That Can Truly Collaborate

The journey to τ2-bench reveals that the next generation of agents must master collaborative problem-solving. The critical, unsolved challenges are:

1. **Consistency:** Achieving reliable performance over many trials (`pass^k).
2. **Rule-Following:** Faithfully adhering to complex, domain-specific policies.
3. **User Guidance:** Effectively communicating with and coordinating an active user in a shared-control environment.

The τ-bench family provides the necessary testbeds to measure what truly matters and build agents ready for the real world.