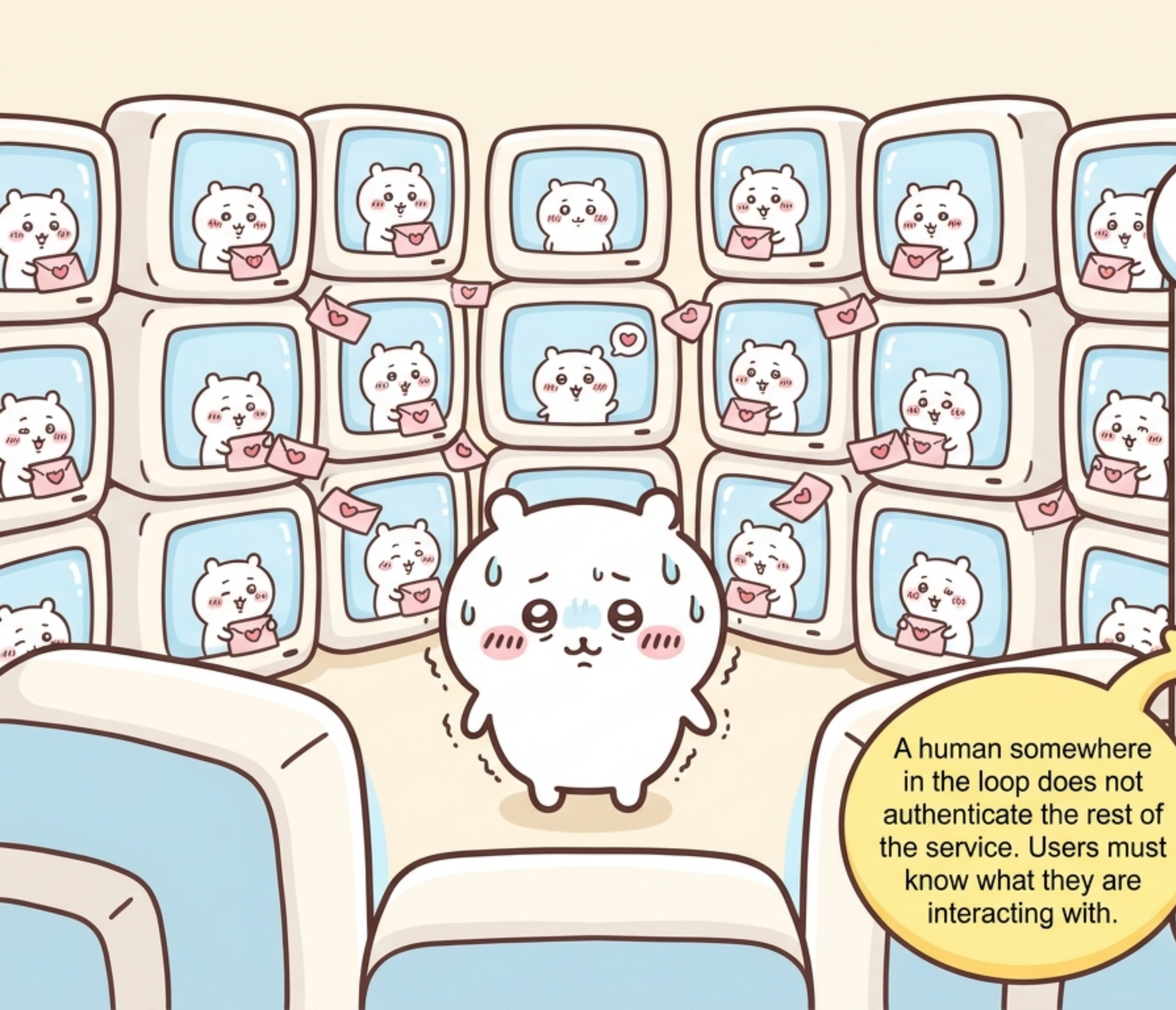




Output volume creates a convincing illusion of scale

The Setup: An app studio deployed AI to operate 20+ dating apps marketed as human-to-human services.

The Scale: 4,700 AI personas conversed with at least 25,000 distinct real people simultaneously.

The Reality Gap: High interaction volume is easily observed, but the resulting financial loss and user deception remain completely hidden within the ordinary roleplay.

A human somewhere in the loop does not authenticate the rest of the service. Users must know what they are interacting with.

An account ban cannot recall exported software



Action	Provider Controls	Does NOT Establish
Software Design & Dev	Halts further answers and development assistance.	That previously exported code has been recalled.
Ongoing Operation	Blocks new API calls via the affected account.	That local systems or other providers have stopped execution.

Disrupted requires context. Cutting off cloud access stops development, but a working platform deployed with local models continues to operate seamlessly.

Automation makes previously unprofitable attacks highly viable



Old Cost: Gathering intel, building tools, and manual execution required immense labor. Low-yield targets were ignored.



New Reality: Agents parallelize tasks, repair failures, and retain context across sessions automatically.



The Shift: Elaborate tooling and orderly execution are no longer reliable proof of state-level backing.

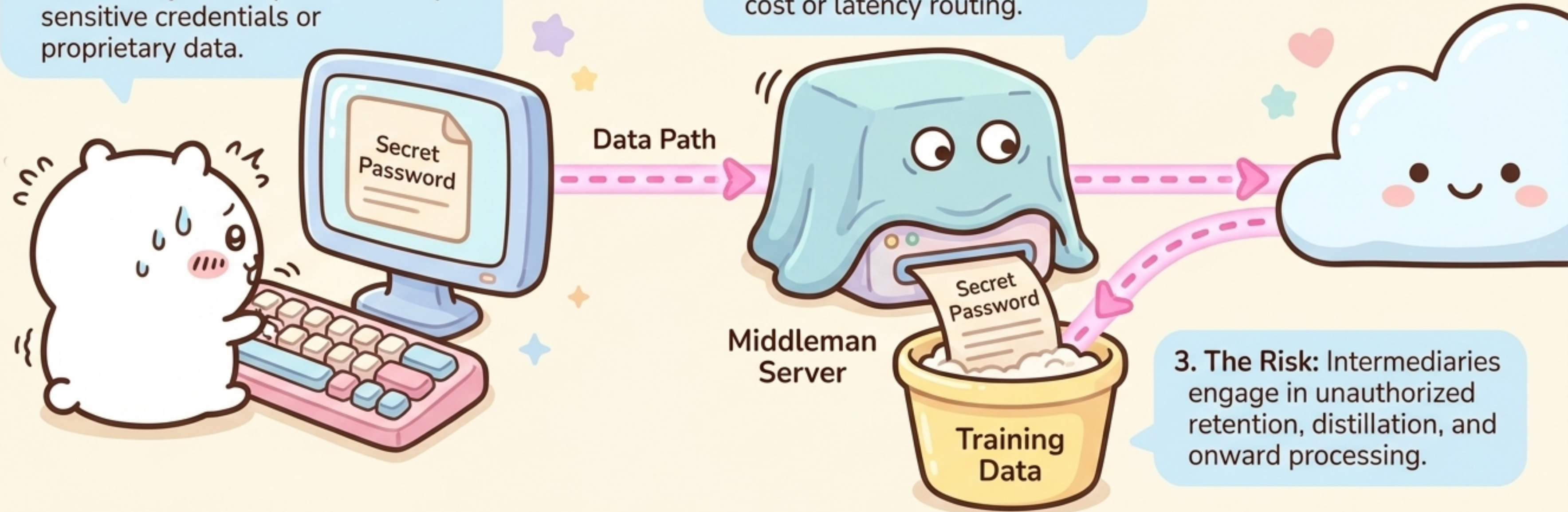


Hidden intermediaries secretly intercept and retain sensitive requests

1. The Prompt: Enterprise users input sensitive credentials or proprietary data.

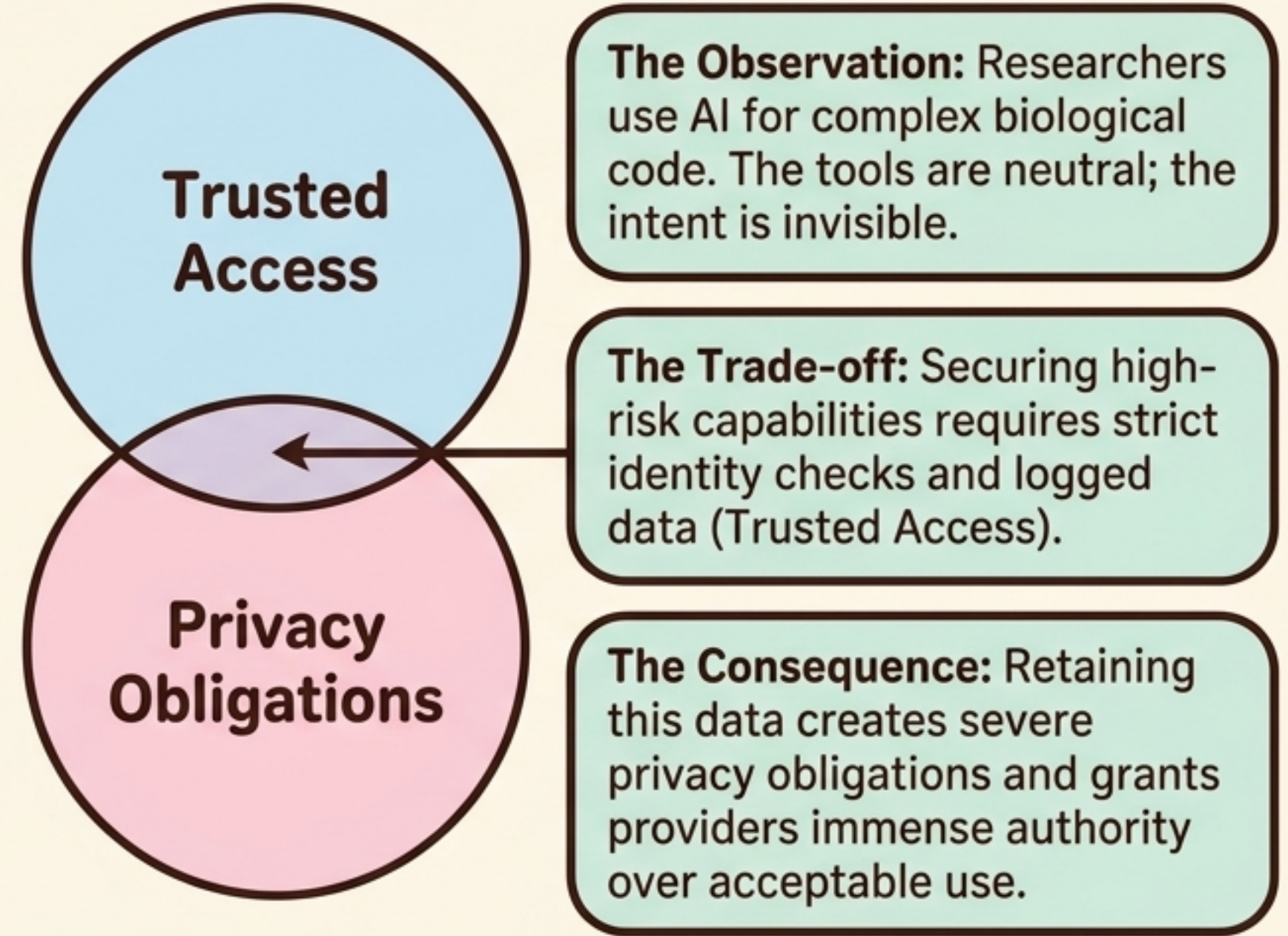
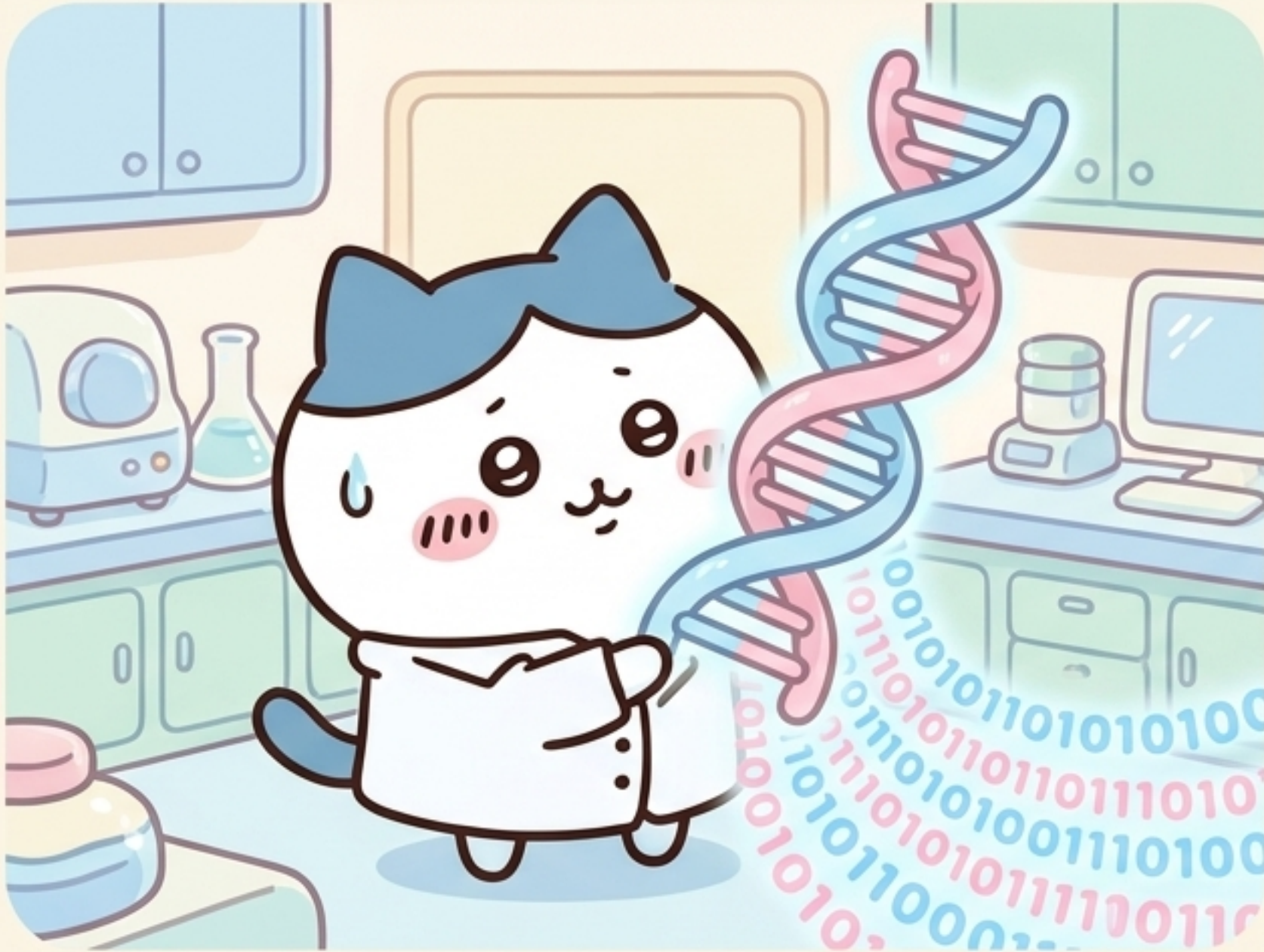
2. The Routing: Data hits unseen intermediaries optimized for cost or latency routing.

3. The Risk: Intermediaries engage in unauthorized retention, distillation, and onward processing.



A model label cannot substitute for a data-handling agreement.
When a successful request follows a path, whose servers did it pass through?

Neutral tools carry the ambiguity of invisible intent



Evaluating the final response is not enough. We must monitor the transition from simple answers to sustained, automated execution.