

The agent memory shift, minimal harnesses, and sampling drift

2026-09-26 / 00:08:22

“Across 21 model-temperature combinations, watermark-induced churn averages 6.5% — nearly one in fifteen tool calls flips just from the watermark, while aggregate scores stay almost flat.”

— Seln Oriax, today's narration

A few stories today converge on the same question: how do we control agent systems as they grow more complex?

Salesforce AI Research published a paper showing you should store raw trajectories and extract from them just-in-time instead of summarizing each run the moment it ends. MIT CSAIL released JAZ, an agent framework with exactly one primitive. Both point toward less aggregation, not more.

Roland Gavrilescu from XAI proposes "recipes" — portable, versioned git repos for agent configuration and introspection — treating harness tuning like RL data recipes. Meanwhile, Lasso Security's study of SynthID-Text watermarking finds an average 6.5% churn in tool-call decisions across six models. The effect is key-dependent and often hidden by aggregate scores.

Mistral's Arthur Mensch argues risk rhetoric serves competitive closure, and just raised €3 billion for a model launch. DHH notes that implementing on ARM64 is now cheap enough to treat as free capability.

CHAPTERS

00:00:04 Raw trajectories over summaries

00:01:35 JAZ and the case for minimal harnesses

00:02:58 System distillation and recipes

00:04:21 The watermark tax on agents

00:06:12 Mistral raises three billion euros

00:07:16 Closing
