VICTOR NDUNDA · INSIGHTS

THE OCTOBER 2026 AI BRIEFING · ESSAY 02

Agents in production — the 2026 field notes

Agents now complete two-thirds of real computer tasks — and still fail one in three. A year of shipping them for paying clients taught me where the real work lives.

OCTOBER 30, 2026 · 6 MIN READ · ESSAY 02 OF 03

This is the year “AI agent” stopped being a demo genre and became a job description. OpenAI’s own adoption study found the fastest growth in agent usage not among developers, but among people in legal, sales, recruiting, and marketing. When your procurement team has agents, the technology is no longer early.

I spent the same year shipping agent systems for paying clients: Keja AI, the real-estate intelligence platform built for CHACADOM INVESTMENTS, now in public trial; Tapi Learn, the multilingual learning platform behind ACLA’s reach to 11,000+ learners in 38 countries; and AgentReplay, my open-source deterministic-replay layer for debugging agents. The view from production is very different from the view from the launch events. Here is what the 2026 hype cycle looks like when it meets real users.

CHAPTER 01

The demo is the easy ten percent

The topline agent number of 2026 sounds like a victory lap: on OSWorld, the benchmark for real tasks across operating systems, agent task success jumped from 12% to roughly 66% in a single year, per Stanford’s 2026 AI Index. The same dataset carries the caveat that decides your architecture: agents still fail about one structured task in three. A demo survives that. A production system does not.

The reason is the asymmetry of consequences. In production, a task done confidently wrong is worse than a task not done at all. On Keja AI, an assistant that mis-states a title-verification status is not a “hallucination” — it is a wrong statement about the largest purchase of someone’s life, wrapped in our interface’s credibility. Every design decision downstream follows from taking that seriously.

So the operating principle I landed on is simple: autonomy is earned per task, per workflow — never granted wholesale. Each function starts sandboxed and supervised; it earns independence by accumulating a measured record on real cases. That is slower to ship and almost impossible to market on a slide. It is also the only pattern I have seen survive contact with month three.

CHAPTER 02

The stack beneath the chat

What actually ships in an agent product is mostly not the chat. It is the unglamorous ninety percent: evaluation harnesses, guardrails, human-in-the-loop review queues, audit trails, and the plumbing to replay incidents. On Keja AI, every listing that enters the 7-stage lifecycle passes document and title verification steps that feed the KEJA Trust Score — 70 listings are already certified at Trust Score ⩾ 80 in public trial, across 12 Kenyan counties, with the assistant answering in English, Swahili, and French. The assistant is the visible tip; the trust pipeline is the product.

Tapi Learn taught the same lesson on a different axis. Its 24/7 AI Career Coach supports learners across 38 countries — 11,000+ of them, rating the platform 4.8/5 (client-reported) — which it can only do because the model is fenced into a coaching scope, with escalation paths for what it must not handle. The coach is bounded, and because it is bounded, it is dependable.

Even my open-source work is really about the ninety percent. AgentReplay records and deterministically replays agent runs with 0.67% overhead and 100% reproduction fidelity, so that when an agent does something bizarre at 2 a.m., you replay it bit-for-bit instead of shrugging. Reliability infrastructure is a product category now — and 2026 is the year that became obvious.

12 → 66%
OSWorld agent task success, one year

1 in 3
structured tasks still fails

0.67%
AgentReplay replay overhead, full fidelity

> The chat is ten percent of the system. The boring ninety percent decides whether it survives month three.

CHAPTER 03

Context is the product

The most under-reported result of the year comes from the State of AI Report: in a controlled coding study, changing only the tools, context, and feedback available to a model lifted its SWE-bench Verified success from 52.5% to 65.5%. No new weights, no fine-tune — thirteen points of real performance from scaffolding alone. Model selection is a lever; the environment you build around the model is a bigger one.

In practice this is what "context engineering" means, and it is where I spend most of my working hours now. On Keja AI, the assistant is only as good as the listing data, verification outputs, and county market context it is grounded in; retrieval is not a garnish, it is the load-bearing wall. The industry is converging on the same insight from the research side: benchmarks saturate in months, but a well-instrumented context pipeline compounds.

The practical consequence for any team: invest in context infrastructure before you invest in another model subscription. A cheaper model with the right data, tools, and feedback loop beats a frontier model guessing in the dark — and the gap is only widening as tooling standards mature.

CHAPTER 04

Agents quietly went mainstream outside engineering

The adoption story of 2026 is not developers vibing with coding agents; it is paralegals, recruiters, and sales teams delegating real work. OpenAI's study found the fastest agent-adoption growth among non-developers. That changes what "good" means in product terms: your users will not read stack traces, so the error surface, the review queue, and the undo button are the experience.

It also changes evaluation. When the user is a hiring manager or a sales lead, “the output looks good” is not a metric — the outcome is. Did the candidate get scheduled? Did the invoice get filed? Task-outcome evals, measured against golden cases from real workflows, are the only acceptance criteria that transfer to this population. Chat-quality benchmarks are nearly irrelevant to them.

There is a second-order effect worth naming: once non-technical teams run agents, the constraint moves to setup — and the State of AI Report locates the whole value gap there (“a human skill issue, not an AI technology issue”). The report’s half-joking call for a “Genius Bar for AI” is, more seriously, the biggest consulting opportunity of the decade.

CHAPTER 05

The economics flipped underneath us

Two opposite things are true at once, and both showed up on my invoices this year. Per-token inference prices keep collapsing while total AI bills keep exploding — because agents make many more calls, carry much more context, and never sleep. Unit costs fall; system costs rise. The teams that win instrument cost per task from day one.

The response is routing discipline: cheap, fast models for routine steps; frontier escalation reserved for genuinely hard hops; hard budgets per workflow. That discipline is now also a defensive requirement, because the release cadence itself is accelerating — Anthropic reported that Claude led 26% of its measured model R&D work by August 2026, up from under 1% in February. When AI is materially building AI, your harness — evals, replay, golden cases — is the only stable ground under the product.

CHAPTER 06

The playbook that survived 2026

01 Define the task and the cost of failure before you pick a model. Most “AI strategy” is really this sentence, unfinished.

02 Build the eval set before the agent — golden cases pulled from real data, not imagined ones.

03 Ship human-in-the-loop first; earn autonomy per workflow as the measured record accumulates.

04 Instrument everything from day one: traces, replays, cost per task. AgentReplay exists because I needed this and nothing did it.

05 Separate interface from policy — you should be able to swap models without rewriting the product.

06 Review unit economics weekly. In a market where prices collapse and bills explode, the drift is silent until it isn’t.

The hype cycle of 2026 will be remembered for benchmark fireworks and bubble debates. From the deployment side, it will be remembered as the year reliability engineering and context infrastructure became the actual product — the year the boring ninety percent got its revenge.

THE OCTOBER 2026 AI BRIEFING · ESSAY 02 OF 3

Three field notes on where AI actually stands.

This essay is part of a three-piece briefing written from Nairobi. The other two:

The State of AI, October 2026
victorndunda.com/blog/articles/state-of-ai-october-2026

The View from Nairobi
victorndunda.com/blog/articles/africa-ai-view-from-nairobi

SOURCES & FURTHER READING

1. **Nathan Benaich — State of AI Report 2026**
   www.stateof.ai
2. **Stanford HAI — 2026 AI Index Report**
   hai.stanford.edu/ai-index/2026-ai-index-report
3. **Keja AI — case study (CHACADOM INVESTMENTS)**
   victorndunda.com/projects/keja-ai/
4. **Tapi Learn — case study (ACLA)**
   victorndunda.com/projects/tapi-learn/
5. **AgentReplay — open source**
   github.com/gadda00/agentreplay

A note on method. External figures trace to the dated sources above — mainly Stanford HAI's 2026 AI Index, the State of AI Report 2026, and October 2026 news reporting. My own client metrics keep their public-trial / client-reported qualifiers. Live links: **victorndunda.com/blog/articles/agents-in-production-2026**

Victor Ndunda

AI Systems Engineer — designing, building, evaluating and maintaining production AI end-to-end. Nairobi, Kenya.

victorndunda.com · github.com/gadda00 · @victor_ndunda

Share this PDF on LinkedIn or X — and tag **@victor_ndunda** so the discussion finds me.

© 2026 Victor Ndunda

NAIROBI · KENYA