AI This Week: The Bill Nobody Costed, From 136.5× Energy Per Query to a December 2 Watermarking Deadline

An agentic query can burn up to 136.5× the GPU energy of a single-turn LLM call. The EU AI Act Omnibus watermarking obligation is due December 2, 2026. Neither number appears in most 2027 budgets.

Agentic AI costs more than the model bill suggests because the model bill is one line of several. Research covered here on September 17 found agentic reasoning burns up to 136.5× more GPU energy per query than a single-turn call. Add licensing scope, pricing curves and three EU compliance regimes, and the gap widens further.

Five posts went up on this site between September 15 and 19, and they were not planned as a series. They ended up as one anyway. The common thread: the cost of running agentic AI in production is being estimated from the wrong inputs. Energy gets estimated from single-call benchmarks. Licensing gets estimated from the license file. Compliance gets

estimated from the headline date rather than the obligations hanging off it. Each time someone measures instead of assumes, the measured number lands outside the range people planned for, and usually in the expensive direction.

The 136.5× figure is a peak, and peaks are what break capacity plans

The research covered on September 17 found that agentic reasoning burns up to 136.5× more GPU energy per query than a single-turn LLM call. That is a peak multiplier, not an average, and the distinction matters for how you read it. The worst case is what blows through a capacity plan. If you sized inference infrastructure on single-call telemetry and then shipped an agent that plans, calls tools, re-reads its own output and retries, you did not size for the workload you deployed. The number is large enough to stop being a footnote in a sustainability report and become a capacity question.

MIT on the QM harness covers the code and stops there

Y Combinator's QM agent harness ships under MIT, about as permissive as open source gets. The MIT license buys you the code and not much else. It says nothing about your model access terms, nothing about what data your agent sends where, and nothing about who carries operational liability when the harness executes something it should not have. Permissive licensing on the orchestration layer is a real gift. It is also routinely mistaken for permissive terms on the whole stack, which is a category error with a legal department attached.

Feature matrices converge, automation pricing models do not

The comparison of n8n, Zapier, Make and OpenAI Agent Builder, published September 18, found the pricing models diverge far more than the feature matrices suggest. Features converge because vendors copy each other. Pricing does not converge, because that is where the margin lives. Two platforms that look interchangeable on a checklist can differ enormously at your particular volume, and the shape of the curve decides whether your

agent workload is cheap at pilot scale and ruinous in production, or the reverse.

Three EU compliance regimes went live inside 48 hours

The EU AI Act became fully applicable on August 2, 2026, with high-risk transition deadlines pushed out to December 2, 2027 and August 2, 2028. The extended high-risk dates are the ones people quote. They are not the ones that bite first.

The one that bites first is December 2, 2026: the Omnibus watermarking obligation. My post on September 15 covered five specific ways that deadline gets misread. Two and a half months from that publication date to the obligation, and the technical work is not a one-line SDK swap for most content pipelines.

Underneath it, two more layers landed in the same week. EU Cyber Resilience Act incident reporting obligations took effect September 11, 2026, per Origin Brief's regulatory weekly, adding a third EU layer alongside the AI Act and the DSA. EU AI Act Article 14 reporting applies from September 11 for certain products already on the market, and Data Act access-by-design from September 12. If your compliance function is one person with a spreadsheet, that week was not survivable in the way people assume compliance weeks are survivable.

This is also why the video piece was framed the way it was. I ranked five AI video models by what an EU studio can legally ship in 2026, using legal shippability rather than benchmark quality, because for a studio operating in the EU that is the binding constraint. The best model you cannot ship is worth less than the fourth-best model you can.

The 10th Circuit answer is a signature, not a ban

The US 10th Circuit Court of Appeals proposed a rule on September 18, 2026 requiring lawyers to certify that AI-prepared filings were independently reviewed for accuracy and citations, per Reuters. If adopted, it takes effect January 1, 2027. The mechanism is worth noting: no ban, no model approval process, just a named human signing that they checked. **That is a liability assignment**, and it is cheap for the court to administer.

Google also shipped in the same week. Gemini 3.8 Live and Gemini 3.8 Live Extended

Thinking launched on September 15, 2026, GA in the Gemini API and AI Studio the same day, rolled out to Search Live globally, and entered private preview in Gemini Enterprise. It is hosted-only. There is no self-hosted option.

That last detail is the one I would put in front of a CTO. For any team whose data residency story depends on running weights inside their own boundary, a hosted-only frontier model is off the table regardless of how good it is. It also feeds the energy question from a different angle: hosted-only means the 136.5× multiplier is Google's electricity problem and your invoice problem, in a currency you cannot audit. You get simplicity and give up visibility. Whether that trade is good depends on whether anyone will ever ask you to account for the energy or the data path.

Meanwhile the release cadence has not slowed. A September roundup counted 50 AI releases in the month, 11 of them in the single week ending September 17.

Models ship weekly, their running cost gets measured yearly

MY TAKE

The shipping rate and the costing rate have come apart. Models ship weekly; the true cost of running them gets measured yearly, if at all. What September showed is that when someone finally measures, the answer is a multiplier rather than a 20% correction. I expect the next two quarters to produce more of these: measured numbers for agentic energy, measured numbers for compliance effort per pipeline, measured numbers for what permissive licensing actually covers. My prediction, and it is only a prediction: the first serious agentic AI budget overrun that goes public will be blamed on infrastructure, and the driver will turn out to be retry loops in agent orchestration that nobody instrumented.

There is a version of this argument that ends in "slow down," and I do not believe it. The pace is fine. The problem is committing to an agentic architecture on the assumption that its running cost is a modest multiple of your current inference bill, then discovering after the contract that the worst case has three digits in front of the decimal point.

I would do three things before the end of this quarter, and none of them require a new

vendor.

Instrument agent energy and token consumption per completed task rather than per call. The 136.5× figure is a peak, and I do not know where any given workload sits on that distribution. Nobody does until it is measured, which is the argument for measuring it rather than waiting for the invoice.

Separate the license audit from the liability audit. MIT on the harness tells you what you may copy. It tells you nothing about model access terms, data exposure, or who is on the hook when the agent acts. Two different reviews, two different people.

Price your automation stack at projected production volume against the actual pricing model rather than the feature list. Four platforms that look alike on features will not look alike on the bill.

And if you publish anything generated or substantially assisted by AI into the EU market, **December 2, 2026 is the date to put on the wall**. Not the 2027 and 2028 high-risk dates. Those are further out and, judging by how often they get quoted in place of the nearer deadline, they are doing real damage as a distraction.

WHAT TO DO WITH THIS

Three deadlines and one multiplier define the next twelve months: December 2, 2026 for EU watermarking, January 1, 2027 if the 10th Circuit rule is adopted, and up to 136.5× GPU energy per agentic query against whatever you budgeted from single-call numbers. If you are sizing an agentic deployment for 2027 and these are not in the model, the model is wrong. Book a call →

WRITTEN BY
Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting →