Breathing Something That Could Have Been Air

On the Opportunity Cost of Generative Compute

Ivor
Independent • Ivorisnoob@gmail.com

July 2026

ABSTRACT

Text-to-video generation is the most energy-intensive inference workload in consumer deployment, and until this month the figures in circulation were guesses. They are no longer. Direct GPU metering across six open models places a single 720p clip between 186 and 1,275 watt-hours depending on architecture and length, with the full tested range spanning 2.95 to 3,364.6 Wh. An eight-second clip costs roughly 1,625 times a single text prompt, and provider-to-provider variation on the identical user-facing task spans one to two orders of magnitude. Over the same period, an audit of the largest video platform found that 21% of short-form videos served to a new account are mass-produced synthetic content, a category sustaining an estimated $117 million in annual advertising revenue.

This essay does not argue that generative models are harmful or that they should be restrained. It argues something narrower: the same class of hardware that produced 214 million predicted protein structures, that found 18 previously unknown vulnerabilities in critical open-source infrastructure in 143 autonomous hours, and whose successor doubled state-of-the-art accuracy on the hardest ligand-binding problems in February 2026, is being pointed at industrial-scale content that nobody requested and nobody remembers. Nothing was stolen. Something was chosen. I set out the current measurements, trace the second-order costs through fraud, attention and recursive data contamination, treat the fungibility objection honestly, and argue that the correct frame is allocation rather than prohibition.

Keywords: generative inference, energy measurement, opportunity cost, synthetic media, compute allocation, model collapse, AI for science

1. INTRODUCTION

I am tired. Not of AI, of what we point it at. We built something that could think about proteins, and we have it making thirty-second videos of a flood that never happened.

That is the whole thesis, and I want to state it at the top so that nobody has to read four sections to find out whether this is another anti-AI essay. It is not. I am not going to argue that the models are bad, that the field is a bubble, that the art is soulless, or that generation should be restricted. Those arguments exist, some of them are good, and none of them are mine. The claim here is about allocation: where finite capacity goes, and who decides.

The distinction matters because the two arguments have different failure modes. The anti-AI argument fails when the technology delivers, and it has delivered. The 2024 Nobel Prize in Chemistry went to Demis Hassabis and John Jumper for protein structure prediction [35]. In February 2026 Isomorphic Labs released IsoDDE, which achieves 50% accuracy on protein-ligand structure prediction for cases with under 20% sequence similarity to training data, against AlphaFold 3's 23.3% [36, 37]. In August 2025, seven autonomous systems in a DARPA competition read 54 million lines of code and handed 18 real, previously unknown vulnerabilities to open-source maintainers [46]. If your position is that this technology is a waste, these results are a refutation, and you lose.

The allocation argument does not have that failure mode, because the successes are its premise. The machines work. That is precisely why it is worth asking what we are running on them.

A note on method. Every number here is sourced, and where sources conflict I say so rather than picking the one that flatters the argument. Where a figure is an estimate rather than a measurement, it is labelled. Where the honest reading weakens my case, I have included it anyway, because a piece complaining about content optimised for impression over accuracy does not get to be that.

2. THE RECEIPT

2.1. The numbers used to be guesses

Until very recently, every public figure for the energy cost of an AI video was reverse-engineered. The most widely circulated one, 3.4 million joules for a five-second clip, came from a May 2025 MIT Technology Review analysis with Hugging Face researchers and was inferred, not metered [4]. Estimates elsewhere ranged from 20 to 100 Wh (Wall Street Journal), 90 Wh (CNET), and 936 Wh (from GPU-rental cost arithmetic), all drawing on the same underlying inference [5, 6]. Sasha Luccioni, one of the Hugging Face researchers behind the original work, objected to her own number's afterlife in the press, arguing that the field should stop reverse-engineering figures from hearsay and start pressuring companies to publish [5].

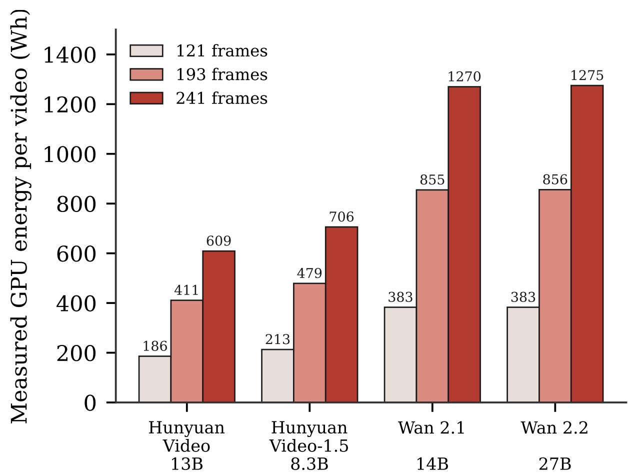

Figure 1. Directly metered GPU energy per generated video at 720×1280, 40 denoising steps, 8×H200, batch size 1. HunyuanVideo-1.5 (8.3B) costs more than HunyuanVideo (13B); Wan 2.1 (14B) roughly doubles HunyuanVideo despite comparable active parameters. Architecture, not model size, sets the bill. Data: [1], Fig. 2.

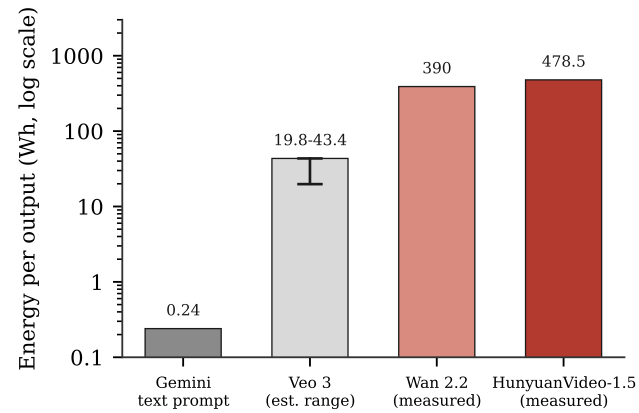

Figure 2. Cost of an eight-second 720p generation, log scale, against a text prompt. Wan 2.2 and HunyuanVideo-1.5 are metered; Veo 3 is estimated via the paper's backward framework and shown as a range, consistent with its reported TPU v6e deployment. The Gemini figure is Google's own disclosure. Sources: [1, 3].

That objection is the right one, and as of this month it has partly been answered.

2.2. Direct measurement, July 2026

Jegham, Gamazaychikov and Luccioni's *Lights, Camera, Carbon*, posted 5 July 2026 and currently under peer review, metered GPU power draw every 100 milliseconds across six open text-to-video models spanning 8.3B to 27B parameters, on three hardware configurations, using pyNVML [1]. These are measurements, not inferences. Figure 1 shows the core result.

Across all tested configurations, energy per video spans nearly three orders of magnitude: from 2.95 Wh for a three-second LTX-2 generation at 1024p on a B200, to 3,364.6 Wh for a twenty-three-second Wan 2.2 generation at 720p with 50 steps on an eight-GPU H200 station [1]. At a fixed five-second, forty-step workload the spread across models is still an order of magnitude: 9.1 Wh for LTX-2, 57.5 Wh for HunyuanVideo-1.5, 113.9 and 114.8 Wh for Wan 2.1 and 2.2, equivalent to running a 1,200 W air fryer for 27 seconds, 2.9 minutes, and 5.7 minutes respectively [1].

Against text, the comparison is stark. An eight-second 720p Wan 2.2 generation on a single H200 draws 390 Wh, which is 1,625 times a Google Gemini text prompt at 0.24 Wh; the same generation from HunyuanVideo-1.5 on eight H200s draws 478.5 Wh, nearly 1,990 times [1, 3]. Figure 2 puts these on one axis.

Three findings matter more than any single number.

Architecture sets the cost, not size. HunyuanVideo-1.5 has 8.3B parameters and consistently burns more than the 13B HunyuanVideo. Wan 2.1 (14B) uses roughly twice HunyuanVideo's energy at comparable active parameter count [1]. You cannot guess a model's footprint from its rumoured size.

Batching buys nothing. Video generation is compute-bound, sustaining 80.9–98.7% of thermal design power throughout inference, so energy per video stays flat across batch sizes: 146.1, 141.3 and 144.2 Wh at batch 1, 2 and 4, a deviation under 3.3% [1]. Unlike text inference, there is no economy of scale. A farm generating a thousand clips pays a thousand times.

The gap between providers is enormous. The authors' framework estimates Veo 3 at 19.8–43.4 Wh for an eight-second 720p clip, and Sora 2.0 Pro at approximately 1,313 Wh for a twelve-second 1080p clip [1, 2]. That is one to two orders of magnitude for what the user experiences as the same product. The waste is not intrinsic to video generation. Some of it is just somebody's architecture.

2.3. At deployment scale

Google's Flow reached 100 million generated videos in 2025 [1]. At the measured rates, 100 million eight-second generations corresponds to 39 GWh for Wan 2.2 and 47.9 GWh for HunyuanVideo-1.5, the annual electricity consumption of roughly 3,645 and 4,477 average US households [1]. The same group's public write-up estimates that Sora 2.0 Pro at a deployment scale of four million users generating two 720p clips a day would consume about 602.6 GWh over six months, comparable to up to 142,000 US households over the same period [2].

I am reporting these as the authors state them. The 100 million figure is a disclosed count; the four-million-user Sora scenario is explicitly hypothetical and should be read that way.

3. TRAJECTORY

The per-clip number is the wrong unit for a policy argument. The right one is the curve, and the curve is documented.

The IEA's *Energy and AI* places global data centre electricity consumption at around 415 TWh in 2024, about 1.5% of world consumption, growing at roughly 12% per year since

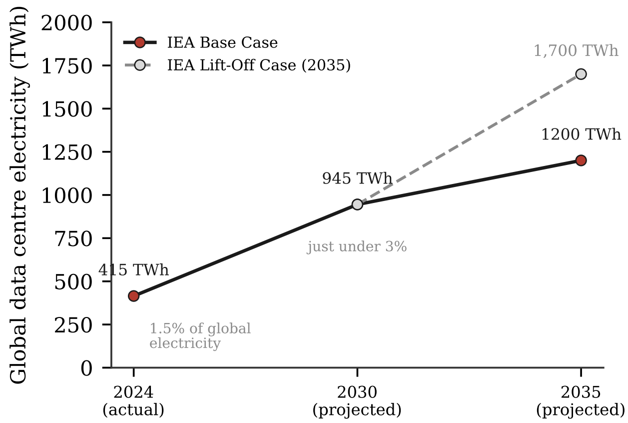

Figure 3. Global data centre electricity consumption, IEA *Energy and AI* Base Case, with the Lift-Off Case for 2035 shown dashed. Data centres are not mostly AI, but accelerated servers account for almost half the net increase to 2030. Source: [7, 8].

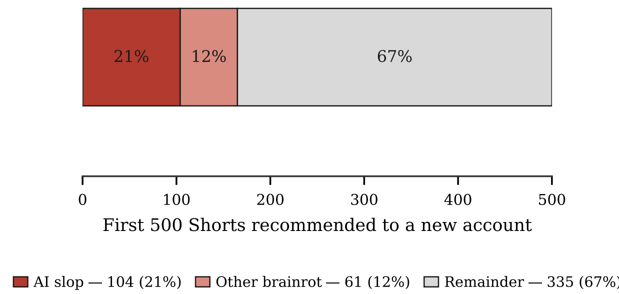

Figure 4. Composition of the first 500 Shorts recommended to a new account: 104 AI slop, 61 other brainrot, 335 remainder. “Other brainrot” is the brainrot category net of the AI-slop subset, to avoid double counting. Source: Kapwing, via [11].

2017, more than four times faster than total electricity consumption [7, 8]. Its Base Case has that more than doubling to around 945 TWh by 2030, just under 3% of global electricity, and rising to roughly 1,200 TWh by 2035; a Lift-Off Case with stronger AI adoption exceeds 1,700 TWh by 2035, around 4.4% of global demand [7, 8]. Figure 3 plots both.

Four things in that report deserve to be quoted accurately rather than sensationally, because the sensational version is wrong in both directions.

It is smaller than the panic suggests. Data centres account for around one-tenth of global electricity demand growth to 2030, less than industrial motors, air conditioning, or electric vehicles [8]. Anyone telling you AI is the main driver of global electricity growth is overstating it.

It is larger than the reassurance suggests. In advanced economies, which have seen decades of essentially flat demand, data centres account for more than 20% of demand growth to 2030 [8]. By the end of the decade the United States is set to consume more electricity for data centres than for the production of aluminium, steel, cement, chemicals and all other energy-intensive goods combined [8]. That is not a rounding error. That is a country deciding what its grid is for.

It concentrates. Nearly half of US data centre capacity sits in five regional clusters [8]. The US and China together account for nearly 80% of global growth to 2030 [7]. Averages hide this: a 3% global share can be a 26% local share.

Inference, not training, is the driver. Electricity consumption in accelerated servers, mainly AI, is projected to grow 30% annually and accounts for almost half the net increase in global data centre consumption [7]. Estimates put inference at 80–90% of AI compute today [9]. Training runs are the number everyone argues about; they are also the one-off. Inference is the weather.

The externality does not stop at the operator. US residential electricity prices rose 42% between 2021 and 2026, with the steepest increases concentrated in high data centre density regions: 94% in Washington, D.C., 74% in Maryland [10]. Virginia’s 2025 rate adjustment added $11 to $13 per month to a typical residential bill [10]. The causal attribution is contested and I am not going to overstate it; rate increases have many drivers. But the IEA independently estimates that around 20% of planned data centre projects are at risk of grid-connection delays [8], which is a physical constraint rather than a rhetorical one.

And efficiency will not save this. Google reduced data centre emissions 12% in 2024 while its absolute data centre electricity consumption grew 27% year-over-year, and its overall greenhouse gas emissions rose 51% between 2019 and 2024 [9]. The IEA’s doubling projection is made *in spite of* anticipated efficiency gains [9]. This is Jevons, and it is old, and it is reliable. Cheaper slop is more slop. The measurement work reinforces it from an unexpected direction: because video generation is compute-bound and batching yields no per-video savings [1], the standard "scale brings efficiency" argument does not even apply here.

4. WHAT THE COMPUTE IS PRODUCING

The energy accounting is only interesting if the output is worth arguing about. So: what is being generated?

The most systematic public audit is Kapwing’s, which created a fresh YouTube account and classified the first 500 Shorts served to it. 104 of them, roughly 21%, were mass-produced synthetic content; a broader “brainrot” category covering repetitive engagement-optimised material accounted for 165 [11]. Figure 4 shows the composition.

This is not marginal. The same study identified 278 channels across the global top-100 rankings consisting entirely of synthetic content, collectively holding 63 billion views, 221 million subscribers, and an estimated $117 million in annual advertising revenue as of October 2025 [11]. YouTube’s CEO named the problem in his January 2026 annual letter; enforcement since has terminated 16 major channels, wiping 4.7 billion views and 35 million subscribers [12]. A March 2026 New York Times investigation found roughly 40% of videos recommended to children, including on the children's product, appear to be synthetic, frequently claiming to be educational while containing misidentified objects and misspelled words [14].

Note what this is not. It is not deception in the classical sense, mostly. It is not a conspiracy. It is not even, in the main, malice. It is an economic gradient. Production cost fell to approximately zero while advertising revenue stayed indifferent to quality, and a system that pays for attention and does not ask about provenance produced exactly what it was configured to produce. Operators run dozens of channels through automation platforms wiring together trend detection, script generation and upload scheduling [13]. Nobody broke a rule. There is no rule.

Gary Marcus, assessing the same phenomenon, called it perhaps the most wasteful use of a computer ever devised [13]. I think that is close, and I think it undershoots in one respect: waste implies an accident. This is the system working.

5. The Second-Order Costs

Energy is the cost I can measure most cleanly, which is exactly why I should not pretend it is the only one or the largest one. Three others are worth setting out, each with its evidential weight stated honestly, because they are unequal.

5.1. Fraud: the strongest evidence

Synthetic media has an obvious criminal application, and unlike the softer harms this one leaves an audit trail. The FBI's IC3 annual report, published April 2026, logged 22,364 US complaints referencing AI with $893.35 million in adjusted losses during 2025 [15, 16]. Total IC3-reported cybercrime losses rose from $16.6 billion in 2024 to $20.9 billion in 2025 [15]. The canonical single incident remains the UK engineering firm Arup, where a finance employee was induced by a deepfaked video conference into making 15 transfers totalling over $25.6 million [17]. Deloitte's Center for Financial Services projects US generative-AI-enabled fraud losses reaching $40 billion by 2027, up from $12.3 billion in 2023, at a 32% compound annual rate [18, 17].

Table 1 sets these out with their measurement basis, because the figures circulating in vendor marketing routinely combine incompatible units into one scary total, and they should not be combined.

Table 1. Second-order cost estimates and what each actually counts. These are not additive. Different bases, different populations, different years.

Figure	What it counts
$893.35M, 22,364 reports (2025) [16]	US IC3 complaints that *referenced* AI. Self-reported; excludes unreported and unrecognised cases.
$20.9B total (2025) [15]	All IC3 cybercrime losses, AI or not. Context, not attribution.
$25.6M (Arup, 2024) [17]	One incident. Illustrative ceiling, not a rate.
$40B by 2027 [18]	Deloitte *projection*, extrapolated by applying fraud-risk scores to 26 IC3 categories. A model, not a count.
7× [19]	Ratio of non-AI "cheap fakes" to AI deepfakes in the 2024 US election cycle (Knight Columbia). Cuts *against* the deepfake panic.

That last row matters. Research reported by Knight Columbia found that cheap fakes made without AI were used roughly seven times more often than AI-generated deepfakes in the 2024 US election cycle [19]. The World Economic Forum ranked AI-driven misinformation the most severe short-term global risk in its 2024 assessment [20], and the honest verdict on the 2024 elections is that this ranking was, at minimum, premature. If you want to argue that synthetic media is corroding democracy, the election evidence is currently weaker than the fraud evidence. I would rather say that than pretend otherwise.

5.2. Attention and mental health: the weakest evidence

This is where I have to be most careful, because it is where the temptation to overclaim is strongest and the data is thinnest.

What exists: a JAMA Network survey of nearly 21,000 US adults in spring 2025 found an association between higher generative AI use and increased risk of depression; about 10% of respondents reported daily use and 5.3% multiple times a day [21]. Gerlich (2025), across 666 participants, found a significant negative relationship between frequent AI use and critical thinking, mediated by cognitive offloading [22]. A BCG study of nearly 1,500 full-time US employees documented what workers call "AI brain fry", acute cognitive fatigue under heavy multi-system AI use [23]. "Brain rot" was Oxford's Word of the Year for 2024; "slop" took the 2025 honours [24, 25].

What that evidence is not: causal. These are cross-sectional surveys and correlational studies. An association between AI use and depression is equally consistent with depressed people using more AI, with both tracking some third variable, or with reverse causation. Nobody has run the experiment that would settle it, and the honest position is that the mental-health case against synthetic media is currently a hypothesis with suggestive correlates, not a finding.

What I will say without hedging is narrower and follows from the design rather than the epidemiology. Infinite scroll was engineered specifically to remove natural stopping points [23]. Synthetic content removes the last production constraint on filling it. Whatever the feed was doing to people before, there is now no upper bound on how much of it there can be. That is a structural claim, and it does not require a p-value.

5.3. Recursive contamination: the most interesting evidence

The third cost is the one the industry has the strongest self-interested reason to care about, which is why it may actually get fixed.

Shumailov and colleagues at Oxford and Cambridge showed

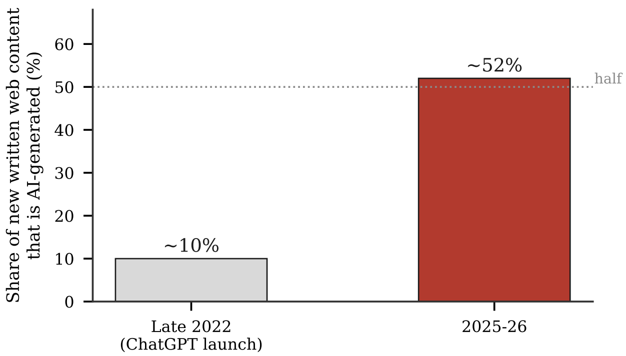

Figure 5. Share of newly published written web content identified as AI-generated, from an analysis of 65,000 articles. A separate analysis of 900,000 pages found 74.2% *contained* AI-generated content, a broader and non-comparable criterion. Source: Graphite and Ahrefs, via [28].

in *Nature* that models trained on recursively generated data collapse: the output distribution narrows, the tails go first, and the model becomes more confident about less [26]. The tails are where rare but crucial patterns live.

This stopped being hypothetical. An analysis by Graphite of 65,000 articles found AI-generated content now accounts for roughly 52% of new written content on the web, up from approximately 10% in late 2022 when ChatGPT launched [28]. A separate Ahrefs analysis of 900,000 newly created web pages in April 2025 found 74.2% contained AI-generated content [28]. Those two figures measure different things, "is AI-generated" versus "contains AI-generated content", and should not be conflated. Figure 5 shows the Graphite series only.

There is a real counter-argument, and it is from within the field. A position paper argues bluntly that model collapse does not mean what most people think it means [27]: the catastrophic results assume each generation *replaces* its predecessor's data, whereas in practice synthetic data accumulates alongside human data, and the collapse largely disappears under accumulation. Frontier labs also curate rather than scrape blindly, and use synthetic data deliberately and successfully.

I find the counter-argument mostly persuasive on the technical claim and irrelevant to the argument here. Even if collapse is avoidable by careful data curation, the curation is a cost, and the cost is imposed on everyone downstream by people who contributed nothing. The commons got harder to use. Somebody pays to clean it. That somebody is not the person who dirtied it. That is the same shape as the electricity bill in the previous section, which is the point.

There is also a symmetry here too neat to skip. Researchers at Texas A&M, UT Austin and Purdue found that continual exposure to short viral social content induces lasting cognitive decline in large language models, including a measurable rise in "thought-skipping", where the model omits parts of its reasoning [29]. We built a machine, pointed it at brainrot, and gave it brainrot. I do not want to lean on the analogy, since a transformer is not a teenager, but the mechanism is worth noticing: feed any pattern-matcher enough low-variance high-stimulation input and its distribution narrows.

6. THE FUNGIBILITY OBJECTION

Here is the strongest argument against everything above, and it deserves to be made properly rather than knocked down.

Compute is not fungible in the way the rhetoric implies. The GPUs rendering a synthetic monkey are not a cancelled cancer trial. No oncologist was told the cluster was busy. The counterfactual in which those cycles go to structural biology is not a real counterfactual, because that demand is bounded by the number of researchers, the availability of validated targets, and the pace of wet-lab confirmation, none of which are compute-limited in the relevant regime. Consumer video generation is a demand that materialised on its own and brought its own capital with it. Some of that capital would not exist at all without that demand. Cutting the slop does not fund the cure. It may simply mean less hardware gets built.

I accept most of this. If the claim were "they stole the cure," the claim would be false, and it should be abandoned rather than defended.

But the objection defeats a claim about physical substitution, and the claim here is about capital and priority. Capacity gets built, powered, cooled, and pointed. Every one of those verbs is a decision made by a person with a budget, and Section 3 establishes that the resulting costs do not stay inside the budget. A share is socialised, through rates, through grid queues, through 20% of planned projects facing connection delays [8]. Section 5 establishes that another share is socialised through fraud losses and through a data commons that now needs cleaning. Once the cost is external, the allocation is not purely a private matter, which means asking about it is not scolding.

There is a second reply, more specific to video. The fungibility objection assumes the waste is intrinsic to the workload. It is not. The measurement shows one to two orders of magnitude between providers on identical output [1, 2]. That gap is not physics. It is engineering effort not spent, and it exists precisely because nobody at the point of use can see it.

7. THE SAME MACHINES, ELSEWHERE

I have spent five sections on the bad allocation. The reason it is infuriating rather than merely disappointing is the good one, so it is worth being specific about what the identical technology does when aimed well, and equally specific about what it has not yet done.

7.1. Structural biology

Determining a protein's three-dimensional structure used to take months or years. AlphaFold reduced it to seconds [31]. The AlphaFold Protein Structure Database went from roughly 365,000 structures at launch in July 2021 to over 200 million a year later, covering nearly every catalogued protein known to science, and 214 million by late 2023 [30, 34]. It is free,

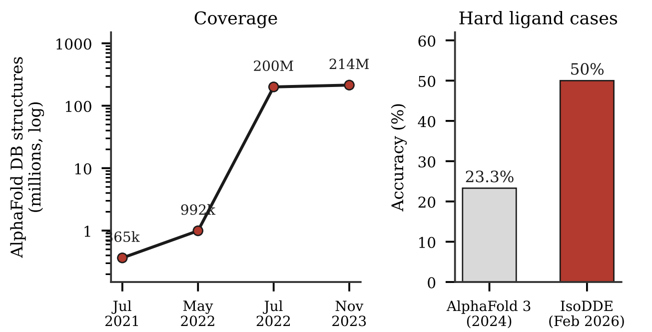

Figure 6. Left: growth of the AlphaFold Protein Structure Database, log scale; the July 2022 step is the release covering nearly the entire catalogued protein universe [30, 31, 34]. Right: accuracy on protein-ligand structure prediction for cases with under 20% sequence similarity to training data [36, 37].

integrated into UniProt, PDB, Ensembl and InterPro, has over three million users across more than 190 countries, and has been cited in more than 35,000 papers [30, 32, 33].

Treat the more expansive counterfactual claims with scepticism. The independent version is more restrained and more convincing: an Innovation Growth Lab analysis found researchers using AlphaFold 2 increased their submission of novel experimental structures by over 40%, and that those structures were more likely to be dissimilar to anything already known [33]. It did not replace the science. It moved the science somewhere it had not been.

The line has not flattened. AlphaFold 3, released May 2024, extended prediction from static protein structures to interactions between proteins, DNA, RNA and small molecules [35]. In February 2026 Isomorphic Labs published IsoDDE, unifying structure prediction, ligand binding, affinity estimation and antibody interaction modelling into one engine and roughly doubling AlphaFold 3's accuracy on the hardest ligand cases, 50% against 23.3% [36, 37]. Figure 6 shows both.

7.2. Drug discovery, and what it has not delivered

The cleanest historical case is halicin. A deep neural network trained on 2,335 molecules learned structural features associated with antibacterial activity against *E. coli* and was applied to a repurposing library of about 6,000 compounds [40, 42]. It surfaced halicin, originally a JNK inhibitor for diabetes, structurally unlike conventional antibiotics, which cleared *C. difficile* infection in a murine model in four days and prevented resistance development in *E. coli* where ciprofloxacin did not [41, 40]. Run against 107 million molecules from ZINC15, the model yielded 23 follow-up candidates of which eight showed antibacterial activity [42]. In 2023 an improved model found an entirely new structural class active against MRSA in two mouse models with low human-cell toxicity [43]; by 2025 the same group had moved from finding molecules to designing them atom by atom, producing leads against multidrug-resistant gonorrhoea and MRSA [44].

Now the honest part, which the field's boosters skip.

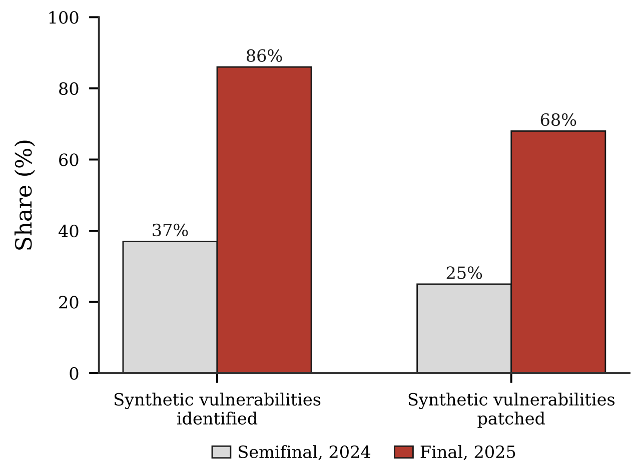

Figure 7. DARPA AI Cyber Challenge, synthetic vulnerability detection and patch rates. Semifinal: DEF CON 32, August 2024. Final: DEF CON 33, August 2025. Sources: [48, 46].

Roughly $60 billion has been invested in AI drug discovery since 2019. Around 175 AI-originated drug programs have entered human trials. None has yet received FDA approval [38]. Isomorphic's first-in-human trial slipped from late 2025 to end of 2026, a delay Hassabis acknowledged at Davos in January [39, 38]. Pharma partnerships involving AI labs carry over $40 billion in cumulative milestone potential, but most of that is deferred and unrealised [37].

So the science side is not a finished story either. It is a story with 175 programs in trials and zero approvals. I include it because a piece that only quoted the good numbers would be doing the thing I am accusing the slop economy of doing.

7.3. Security: the cleanest case available

If drug discovery is the most promising allocation and slop the worst, autonomous security research is the one where the ledger already balances, and it is the least discussed.

DARPA's AI Cyber Challenge ran 2023–2025 with ARPA-H, $29.5 million in prizes, and participation from Anthropic, Google, Microsoft and OpenAI [47, 45]. In the final at DEF CON 33, seven teams ran fully autonomous Cyber Reasoning Systems for roughly 143 hours across 53 challenge projects derived from real critical-infrastructure software, totalling 54 million lines of code. Each team was capped at $85,000 in cloud compute and $50,000 in LLM API credits [45].

The systems discovered 54 unique synthetic vulnerabilities and patched 43. More importantly, because the challenges were built on real software, they surfaced 18 additional genuine, previously unknown vulnerabilities, which were disclosed to open-source maintainers, along with 11 patches for real flaws [46]. Figure 7 shows the year-over-year jump from the 2024 semifinal.

Do the arithmetic that this essay has been doing throughout. Roughly $135,000 of compute and API credits per team, 143

Table 2. Why some allocations compound and others evaporate.

	Good allocation	Slop
Search space	Too large for humans (107M molecules; 54M lines of code)	Small; a prompt
Evaluator	Cheap, approximate, principled	Engagement
Verification	Expensive, external, real (wet lab; maintainer review)	None
Artefact	Durable, checkable, outlives the compute	Half-life of a scroll
Who benefits	Anyone, later, free	The uploader, once

hours, and the output is 18 real vulnerabilities in the software underpinning critical infrastructure, plus every finalist system open-sourced [45, 48]. Set that against 278 channels earning $117 million a year [11]. Same industry, same silicon, same decade.

Google's Big Sleep is the individual-case version: an agent from DeepMind and Project Zero that found its first real-world vulnerability in November 2024, and went on to discover CVE-2025-6965 in SQLite, a critical flaw known only to threat actors and at risk of active exploitation [49, 50]. It has since reported twenty previously unknown vulnerabilities in widely used open-source software including FFmpeg and ImageMagick [51]. The 2026 International AI Safety Report singles out vulnerability discovery as an area with particularly strong evidence of meaningful AI assistance [52].

The honest caveat, again. The same capability cuts both ways: Veracode's 2025 GenAI Code Security Report found 45% of AI-generated code samples introduced OWASP Top 10 vulnerabilities [50]. AI is writing insecure code faster than it is fixing it, and the finding that models are good at discovering vulnerabilities is not unambiguously good news, since attackers read the same papers.

7.4. The general pattern

The successful cases share a structure, set out in Table 2. A search space too large for humans. A cheap approximate evaluator. An expensive verification step the model does not replace but rather aims. And a durable artefact at the end that outlives the compute and can be checked by someone who was not there.

Slop inverts every row. Four hundred and seventy-eight watt-hours becomes eight seconds of attention and then nothing at all. The energy is not recoverable, and neither is the attention.

8. DISCUSSION

8.1. This is not a prohibition argument

I want to close off the reading where this becomes a call to restrict video generation, because that reading is both wrong and a practical dead end. Restriction is unenforceable at the model layer: a model cannot know what a video is for, and the request looks identical either way. It is under-inclusive, since the same clip is fine as a storyboard and corrosive as a fake outbreak. And it is a category error. The problem is not that generation is possible. It is that generation is free and distribution is frictionless.

8.2. Where the friction should go

If the diagnosis is an economic gradient, the intervention belongs where the money is, not where the model is.

Energy disclosure should be mandatory, not archaeological. The most damning fact in this essay is that the best available figures for a mass-market consumer product had to be recovered by researchers metering open substitutes and fitting scaling laws to API latency [1]. Google published a median per-prompt figure for text [3], which proves it is possible and makes the silence elsewhere a choice.

Publish the architecture penalty. The single most useful finding in the measurement literature is the one-to-two-order-of-magnitude gap between providers for identical user-facing output [1]. If that gap were visible at point of use, it would move demand. It is invisible by default, and its invisibility is worth money to the least efficient provider.

Monetisation is the actual lever. The enforcement that worked, demonetisation and termination of high-volume synthetic channels [12], worked because it attacked the gradient rather than the capability. $117 million in annual revenue [11] is the entire explanation for the phenomenon. Remove it and the compute reallocates itself without anybody legislating a model.

Provenance has a deadline, and it is now. The EU AI Act's content labelling obligations take effect in August 2026, with penalties up to EUR 35 million or 7% of global turnover; 47 US states have enacted 169 deepfake laws since 2022 [19]. This is the first real test of whether provenance can be structural rather than voluntary. Voluntary disclosure fails for precisely the population it needs to catch.

8.3. Limitations

The energy figures for open models are metered and I treat them as solid; the proprietary figures for Veo 3 and Sora 2.0 Pro are estimates from a framework whose paper is under review, and are labelled as such throughout. The measurements count GPU energy only, excluding cooling, networking and embodied carbon, and so understate total footprint by an unknown factor. The IEA projections are scenarios, not forecasts, and the agency says so. The slop-prevalence figures come from a commercial platform's audit of a single service, and "AI slop" is a human judgement against a fuzzy standard. The fraud figures come substantially from vendors with a product to sell and count incompatible things, which is why Table 1 exists. The mental-health evidence is correlational and I have not treated it as more. The attribution of residential rate increases to AI workloads specifically is contested. And the counterfactual at the centre of this essay, the world where the capacity went elsewhere, is not observable, which is exactly why I have argued about priority rather than substitution.

9. CONCLUSION

Here is the receipt, plainly, with the current numbers. An eight-second 720p clip metered at 390 to 478.5 watt-hours, between 1,625 and 1,990 times a text prompt, with no batching discount and a hundredfold spread between providers who will not tell you which one they are [1]. Roughly one in five short videos served to a new account is that [11]. The channels producing them hold 63 billion views and $117 million a year [11]. Data centre electricity more than doubles to 945 TWh by 2030, and by the end of the decade the United States spends more power on data centres than on aluminium, steel, cement and chemicals combined [8].

Against that: $135,000 and 143 hours found 18 real vulnerabilities in the software the world runs on [45, 46]. 214 million protein structures, free, in 190 countries [30, 32]. And five months ago, the state of the art on the hardest problem in ligand binding doubled [36].

Nothing was stolen. That is the part that gets me. There was no cure sitting in a queue that got bumped. There was a choice, made over and over, at enormous scale, by people who could have pointed the machine anywhere.

It could have been air. It could have been the thing running quietly under a hospital that nobody notices because it works. Instead we are breathing this.

REFERENCES

[1] N. Jegham, B. Gamazaychikov, S. Luccioni, “Lights, Camera, Carbon: Architectural Scaling Laws for Video Generation Energy Consumption,” arXiv:2607.04553 [cs.MM], 5 July 2026. Under peer review. https://arxiv.org/abs/2607.04553

[2] Sustainable AI Group, “Lights, Camera, Carbon,” 7 July 2026. https://sustainableaigroup.com/lightscameracarbon

[3] C. Elsworth et al., “Measuring the Environmental Impact of Delivering AI at Google Scale,” arXiv:2508.15734, 2025.

[4] MIT Technology Review, “We did the math on AI’s energy footprint,” May 2025. https://www.technologyreview.com/2025/05/20/1116327/ai-energy-usage-climate-footprint-big-tech/

[5] Gizmodo, “All the Bad Things That Can Happen When You Generate a Sora Video,” November 2025.

[6] A. Alexander, “Every Sora AI video burns 1 Kilowatt hour,” Reclaimed Systems, November 2025.

[7] International Energy Agency, “Energy demand from AI,” in *Energy and AI*, April 2025. https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai

[8] International Energy Agency, “Executive summary,” in *Energy and AI*, April 2025. https://www.iea.org/reports/energy-and-ai/executive-summary

[9] Brookings Institution, “Global energy demands within the AI regulatory landscape,” April 2026. https://www.brookings.edu/articles/global-energy-demands-within-the-ai-regulatory-landscape/

[10] Tech Times, “AI Energy Consumption Is Raising Household Bills,” June 2026.

[11] Search Engine Journal, “YouTube’s AI Slop Problem And How Marketers Can Compete,” March 2026, reporting the Kapwing AI Slop Report. https://www.searchenginejournal.com/youtubes-ai-slop-problem-and-how-marketers-can-compete/567297/

[12] OutlierKit, “YouTube’s AI Slop Crackdown: 4.7 Billion Views Wiped,” 2026.

[13] TokenRing, “The ‘AI Slop’ Crisis: 21% of YouTube Recommendations are Now AI-Generated,” January 2026.

[14] Wikipedia, “AI slop,” accessed July 2026, reporting the March 2026 New York Times investigation. https://en.wikipedia.org/wiki/AI_slop

[15] The World Data, “AI Fraud Statistics 2026,” April 2026, compiling FBI IC3 2025 Annual Report figures. https://theworlddata.com/ai-fraud-statistics/

[16] Memeburn, “Deepfake Statistics 2026,” July 2026, citing the FBI IC3 report published April 2026.

[17] TruthScan, “Deepfake Statistics 2026: AI Fraud Data and Trends,” 2026, compiling Deloitte and Arup incident figures.

[18] Adaptive Security, “Deepfake Statistics 2026,” 2026, citing Deloitte Center for Financial Services, *Deepfake Banking Fraud Risk on the Rise*, 2024.

[19] StationX, “Deepfake Statistics 2026: Growth, Fraud & Detection Data,” July 2026, citing Knight Columbia and MultiState. https://app.stationx.net/articles/deepfake-statistics

[20] Stingr.ai, “Deepfake Statistics 2026 (Verified Data),” June 2026, citing WEF *Global Risks Report* 2024 and FBI IC3.

[21] Blue Cross Blue Shield of Michigan, “How Does AI Affect Your Brain?” February 2026, reporting a JAMA Network survey of nearly 21,000 US adults.

[22] “AI Empathy Erodes Cognitive Autonomy in Younger Users,” arXiv:2603.29886, 2026, citing Gerlich (2025).

[23] George Mason University College of Public Health, “AI and the Rise of Cognitive Overload,” March 2026, reporting Bedard et al. (2026), BCG/HBR. https://publichealth.gmu.edu/news/2026-03/ai-and-rise-cognitive-overload

[24] Oxford University Press, Word of the Year 2024: “brain rot”.

[25] “AI generates well-liked but templatic empathic responses,” arXiv:2604.08479, 2026, noting “slop” as 2025 word of the year.

[26] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, Y. Gal, “AI models collapse when trained on recursively generated data,” *Nature*, 2024. Author Correction DOI: 10.1038/s41586-025-08905-3

[27] “Position: Model Collapse Does Not Mean What You Think,” arXiv:2503.03150, 2025.

[28] Forbes Councils, “Synthetic Data & The Quiet Feedback Loop That Will Degrade AI Over Time,” June 2026, reporting Graphite and Ahrefs analyses.

[29] Fortune, “Just like humans, AI can get ‘brain rot’ from low-quality text,” October 2025, reporting Texas A&M, UT Austin and Purdue preprint.

[30] M. Varadi et al., “AlphaFold Protein Structure Database in 2024,” *Nucleic Acids Research*, 52(D1):D368, 2024. DOI: 10.1093/nar/gkad1011

[31] Google DeepMind, “AlphaFold reveals the structure of the protein universe.”

[32] Google DeepMind, “AlphaFold.” https://deepmind.google/science/alphafold/

[33] Google DeepMind, “AlphaFold: Five Years of Impact,” November 2025, reporting Innovation Growth Lab analysis.

[34] Wikipedia, “AlphaFold,” accessed July 2026.

[35] FinancialContent, “The Biological Singularity: How AlphaFold 3 Is Rewriting the Blueprint of Drug Discovery,” February 2026.

[36] DeepCeutix, “Google Just Built the Best Drug Design AI in History. You Can’t Use It,” March 2026, on the IsoDDE technical report of 10 February 2026.

[37] Presenc AI, "AI Drug Discovery: IsoDDE and AlphaFold in 2026," May 2026.

[38] AIM Media House, "2026 Is the Year AI Drug Discovery Meets Clinical Reality," July 2026.

[39] IntuitionLabs, "Isomorphic Labs & AlphaFold: AI Drug Discovery in Trials," 2026.

[40] J. M. Stokes et al., "A Deep Learning Approach to Antibiotic Discovery," *Cell*, 180(4):688–702, 2020. PMID: 32084340

[41] A. Trafton, "Artificial intelligence yields new antibiotic," MIT News, 2020.

[42] Chemistry World, "AI tool screens 107 million molecules, discovers potent new antibiotics."

[43] MIT News, "Using AI, MIT researchers identify a new class of antibiotic candidates," December 2023.

[44] Fierce Biotech, "Generative AI models build new antibiotics starting from a single atom," August 2025.

[45] "SoK: DARPA's AI Cyber Challenge (AIxCC): Competition Design, Architectures, and Lessons Learned," arXiv:2602.07666, February 2026. `https://arxiv.org/abs/2602.07666`

[46] Dark Reading, "DARPA: Closing the Open Source Security Gap With AI," August 2025. `https://www.darkreading.com/cloud-security/darpa-closing-open-source-security-gap-ai`

[47] "ATLANTIS: AI-driven Threat Localization, Analysis, and Triage Intelligence System," arXiv:2509.14589, 2025.

[48] AppSec Santa, "AI Pentesting Agents 2026," June 2026.

[49] Google, "A summer of security: empowering cyber defenders with AI," 2025. `https://blog.google/innovation-and-ai/technology/safety-security/cybersecurity-updates-summer-2025/`

[50] Keysight, "Security Highlight: AIxCC 2025," August 2025, citing Veracode's 2025 GenAI Code Security Report.

[51] Techzine, "Google's AI Big Sleep discovers twenty new security vulnerabilities in open source," August 2025.

[52] *International AI Safety Report 2026*, arXiv:2602.21012.