Will 2026 see the end of the AI Hype?

5 Hype Traps We Need to Ignore to Move Forward

Dylan Seychell Follow 16 min read · Jan 4, 2026

TL;DR: The “Magic” phase of AI is over. In 2026, the industry is waking up to a harsh hangover: thousands of “wrapper” startups are dying, and agentic workflows are hitting the wall of physics and cost. This article argues that the future isn’t about bigger models or more intelligent agents, but about metrics, energy efficiency, and local deployment. If you want to survive the correction, stop chasing the hype and start counting the watts.

Key Takeaways:

- **The Market Correction:** With 95% of pilots failing to deliver ROI, the “wrapper” economy is imploding. 2026 is the end of features masquerading as businesses.
- **The Reality Check:** Progress is hitting two walls: **Engineering limits** (like agentic “Loops of Death”) and **Cultural resistance** (the “Human Firewall” in healthcare and education).
- **The 2026 Pivot:** The industry is shifting from massive cloud models to **Small Language Models (SLMs)** and **Edge AI**. Success now depends on vertical integration and energy efficiency — not magic.

. . .

Introduction

In 2025, MIT’s Project NANDA reported that 95% of generative AI pilots/enterprise initiatives fail to deliver measurable ROI or P&L impact. During the same time, Gartner also noted that worldwide generative AI spending reaches $644 billion, a 76.4% YoY surge. It also reported a trend similar to MIT's, in which application-layer revenue lags far behind infrastructure costs.

This is a market correction we hear very little about, despite the constant hype.

After three years of predictions about Agentic AI, AGI whispers, and sector transformation, the sobering 2025 statistics tell a different story. To begin, one needs to carefully examine the source of information to distinguish signal from noise. Generally, a good indicator is whether the source is someone who *builds* or someone who merely *discusses* AI.

- **The 2026 Pivot:** The industry is shifting from massive cloud models to **Small Language Models (SLMs)** and **Edge AI**. Success now depends on vertical integration and energy efficiency — not magic.

. . .

Introduction

In 2025, MIT’s Project NANDA reported that 95% of generative AI pilots/enterprise initiatives fail to deliver measurable ROI or P&L impact. During the same time, Gartner also noted that worldwide generative AI spending reaches $644 billion, a 76.4% YoY surge. It also reported a trend similar to MIT's, in which application-layer revenue lags far behind infrastructure costs.

This is a market correction we hear very little about, despite the constant hype.

After three years of predictions about Agentic AI, AGI whispers, and sector transformation, the sobering 2025 statistics tell a different story. To begin, one needs to carefully examine the source of information to distinguish signal from noise. Generally, a good indicator is whether the source is someone who *builds* or someone who merely *discusses* AI.

- **The 2026 Pivot:** The industry is shifting from massive cloud models to **Small Language Models (SLMs)** and **Edge AI**. Success now depends on

vertical integration and energy efficiency — not magic.

. . .

Introduction

In 2025, MIT's Project NANDA reported that 95% of generative AI pilots/enterprise initiatives fail to deliver measurable ROI or P&L impact. During the same time, Gartner also noted that worldwide generative AI spending reaches $644 billion, a 76.4% YoY surge. It also reported a trend similar to MIT's, in which application-layer revenue lags far behind infrastructure costs.

This is a market correction we hear very little about, despite the constant hype.

After three years of predictions about Agentic AI, AGI whispers, and sector transformation, the sobering 2025 statistics tell a different story. To begin, one needs to carefully examine the source of information to distinguish signal from noise. Generally, a good indicator is whether the source is someone who *builds* or someone who merely *discusses* AI.

- **The 2026 Pivot:** The industry is shifting from massive cloud models to **Small Language Models (SLMs)** and **Edge AI**. Success now depends on vertical integration and energy efficiency — not magic.

. . .

Introduction

In 2025, MIT's Project NANDA reported that 95% of generative AI pilots/enterprise initiatives fail to deliver measurable ROI or P&L impact. During the same time, Gartner also noted that worldwide generative AI spending reaches $644 billion, a 76.4% YoY surge. It also reported a trend similar to MIT's, in which application-layer revenue lags far behind infrastructure costs.

This is a market correction we hear very little about, despite the constant hype.

After three years of predictions about Agentic AI, AGI whispers, and sector transformation, the sobering 2025 statistics tell a different story. To begin, one needs to carefully examine the source of information to distinguish signal from noise. Generally, a good indicator is whether the source is someone who *builds* or someone who merely *discusses* AI.

- **The 2026 Pivot:** The industry is shifting from massive cloud models to **Small Language Models (SLMs)** and **Edge AI**. Success now depends on vertical integration and energy efficiency — not magic.

. . .

Introduction

In 2025, MIT's Project NANDA reported that 95% of generative AI pilots/enterprise initiatives fail to deliver measurable ROI or P&L impact. During the same time, Gartner also noted that worldwide generative AI spending reaches $644 billion, a 76.4% YoY surge. It also reported a trend similar to MIT's, in which application-layer revenue lags far behind infrastructure costs.

This is a market correction we hear very little about, despite the constant hype.

After three years of predictions about Agentic AI, AGI whispers, and sector transformation, the sobering 2025 statistics tell a different story. To begin, one needs to carefully examine the source of information to distinguish signal from noise. Generally, a good indicator is whether the source is someone who *builds* or someone who merely *discusses* AI.

building this and our other solutions, we could have built it as a SaaS wrapper around OpenAI's or Google's vision APIs. Instead, we trained custom computer-vision models, published our work in peer-reviewed IEEE conferences, and advanced the technology from TRL 3 to TRL 7. The difference is that we at the University of Malta own the IP, control the costs, and can deploy without internet connectivity.

A sample from our SODA Dataset used in our aerial litter detection projects, showing how the same item of litter is visible from different altitudes.

A wrapper competitor could appear tomorrow and undercut us on price using a platform's latest vision capabilities. However, they can't match our domain specificity and would entail significant technical debt. For example, detecting litter types at altitudes of 1–30 metres across Maltese terrain, with material classification to optimise recycling. That specificity is defensible. Generic wrappers are not.

This wrapper collapse sets the stage for deeper challenges in 'agentic' AI hype.

2. Technical Failures: The 'Agentic' Reality

"Agentic AI" has become the industry's buzzword of the moment. It is marketed as autonomous digital colleagues capable of complex workflows. This has indeed been the hope and vision of AI for at least the past 20 years. The promise was evolution from "copilots" (who assist) to "autopilots" (who take over).

2.1 Engineering Challenges

The technical reality has failed to meet these expectations. Gartner predicts that over 40% of agentic AI projects will be cancelled by the end of 2027. This is not because they are rebranded RPA, but because genuine agentic systems exhibit specific, documented failure modes.

The failures and challenges of Agentic AI extend beyond the confabulations and hallucinations present in Generative AI. They are generally organised accordingly:

1. The **Loop of Death** (or recursive reasoning failure) occurs when agents enter infinite loops, repeatedly planning without execution or endlessly refining queries without convergence. This failure mode arises when the reward function overweights thoroughness relative to task completion, leading to algorithmic paralysis. In autonomous systems that require real-world interaction, this represents a fundamental challenge: optimising for confidence scores that approach certainty prevents action in stochastic environments where perfect information is unattainable.
2. As conversations and tasks increase, **Context Window Overflow** occurs because memory becomes saturated with noise that cannot be readily avoided in an agentic context. Despite the marketing of "million-token context windows," agents still experience context drift. Performance degrades non-linearly; for example, an agent that is 95% accurate at step one may be only 60% accurate at step ten.
3. Corporate agents typically lack persistent memory, resulting in a **Learning Gap**. They do not learn from feedback. If you correct an error today, the agent will likely make the same error tomorrow because the learning is not written back to the model's weights or a vector store.

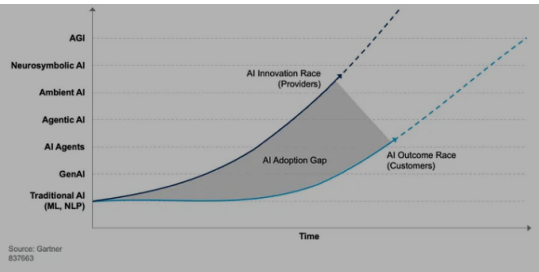
These are engineering challenges that will be overcome sooner rather than later, but they still significantly affect effectiveness and trust. This technical debt undermines trust when valuable topics such as Agentic AI are overhyped, leaving decision-makers disillusioned.

2.2 Agent Washing

The intense demand for agentic capabilities has led to widespread "agent washing." Some vendors are rebranding legacy chatbots, rule-based RPA scripts, and simple automation tools as "Autonomous AI Agents" to capture investor and market interest.

Most implementations are precisely what we've been building for twenty years, just with a new coat of fresh and trendy paint. It is essentially an LLM chained to APIs with a planning loop on top. This does not mean that the giants are not building refined agentic workflows, but most of the noise we hear about Agentic AI is often associated with chained third-party LLMs.

Gartner estimates that, among thousands of vendors claiming agentic capabilities, only about 130 offer genuine autonomous agent technology capable of dynamic planning and adaptability. This market noise complicates vendor selection because buyers struggle to distinguish genuine innovation from rebranded legacy software.



"The mass proliferation of AI providers launching agentic models, agentic-integrated platforms and other agent-infused products far exceeds the present demand" (Gartner, October 2025)

In structured environments, these agents deliver value. However, in messy real-world scenarios, they only amplify human and organisational shortcomings. Moreover, they still hallucinate, drift off-task, or fail spectacularly on edge cases. Guardrails help, but they are brittle and expensive to maintain.

2.3 Underestimated Costs

Agents operate on loops. To complete a single objective, such as "research this competitor," an agent might trigger dozens of internal reasoning steps and external API calls. Each step consumes tokens. A task that costs a human fractions of a cent can cost an agent significantly more if it gets stuck in loops or pursues a hallucination.

Many organisations applied agents to deterministic tasks that are better suited to traditional automation. This is an expensive misuse of technology. Using a probabilistic LLM (or any AI for that matter) for a deterministic task is never followed by a favourable economic outcome.

The reality for 2026 may not be a workforce of digital employees. I rather suspect a wave of quiet cancellations. Enterprises are already shelving overambitious agent projects simply because the ROI does not add up as suggested in the reports cited above

3. Cultural Barriers: The Human Firewall of Resistance

I have read articles and heard keynotes predicting that Education and Healthcare (among others) will experience broad adoption of AI in 2026. The reality I have been experiencing on the ground is more stubborn. While, as builders, we look to benchmarks and metrics, these collide with distinct cultures, ethics, and operational realities across specific industries. Applying AI and Machine Learning to a particular domain is a beast of its own.

The first thing you have to learn in business is that you cannot change a culture. If the culture flexes, it wipes you out, so you can only nudge it gently and gradually. From the various AI projects I have built or managed, I can attest to this.

3.1 Healthcare

A 2025 study from Johns Hopkins University uncovered the "Competence Penalty". Physicians who rely on generative AI for diagnostic decision-making are perceived by their peers as less competent and less skilled than those who rely on their own judgment or traditional tools.

Medical culture is built on apprenticeship, pattern recognition developed over thousands of clinical hours, and the social performance of expertise. An AI-assisted diagnosis, even if statistically superior, undermines the physician's authority in the eyes of colleagues.

The most effective adoption remains in low-stakes administrative tasks such as ambient notes during consultations, insurance pre-authorisation drafting, and appointment scheduling. Core diagnostic and treatment decisions remain largely untouched. Physicians cite immature tools, liability fears, and deep unease about losing patient contact. They ultimately protect domains in which human judgment remains constitutive of trust.

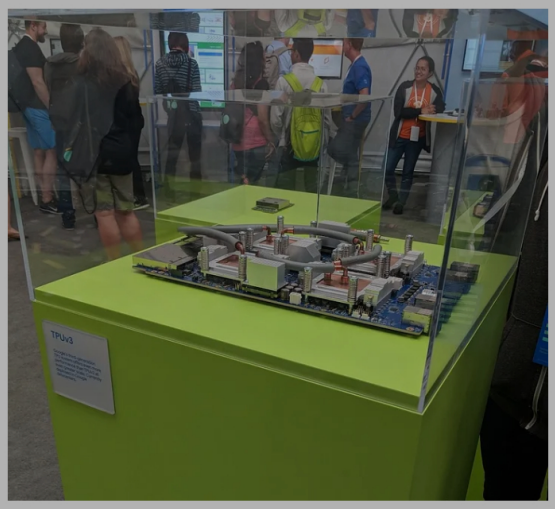
assessment methods in education so that we help our students improve the processes they undertake, rather than relying on rigid measures of progress. In 2026, I will be publishing a book that explores this in depth.

These and other professionals across sectors are safeguarding domains in which human judgment and relationships remain profoundly important. The "revolution" we hear about amid the hype will more closely resemble cautious, incremental pilots within a brave will to change.

4. Emerging Solution: The Energy Efficiency Imperative

The issue of AI and Energy Consumption is often treated as an abstract moral failing. This misses a fundamental economic reality of AI. Companies spending billions on compute infrastructure are not doing it because they enjoy burning money. They're trapped in a cost structure they desperately want to escape.

The AI giants have strong financial incentives to make their models smaller, faster, and cheaper to run. Inference costs eat through their margins. Google's investment in custom Tensor Processing Units (TPUs) rather than relying solely on NVIDIA GPUs demonstrates this economic imperative. Purpose-built silicon optimised for neural network operations delivers superior performance per watt, thereby directly reducing operational costs at the scale of billions of daily inferences. Moreover, TPUs provide Google greater control over the entire AI stack. I wrote an article on Google's AI strategy that provides additional context.



A picture I took of a TPUv3 on display during Google's IO of 2018. The (smaller) TPUv1 can be seen in the other display in the background. For context, Google's TPUs are currently at v7, which is explicitly designed for inference.

In reality, 2025 marked a shift toward efficiency, with smaller language models (SLMs) delivering comparable results with fractions of the power, enabling local deployment on devices without constant reliance on the cloud.

> *Fun fact: Before 2018, Google Translate primarily relied on cloud-based neural machine translation, but in 2018, Google introduced compressed neural models for offline, on-device use across 59 languages on Android and iOS, delivering remarkable scale with significant savings in data usage, latency, and accessibility worldwide.*

Just as "agent washing" rebrands legacy tools, 2025 saw "efficiency washing," in which vendors hyped modestly compressed models as revolutionary SLMs

in which vendors hyped modestly compressed models as revolutionary SLMs to secure funding. Actual efficiency requires purpose-built designs such as Meta's Llama 4 Scout or xAI's Grok-3 Mini, which balance size with reasoning capability.

This is not charity or altruism. These companies are motivated to address this issue to optimise profitability, monetise through paid access to updates and premium features, and reduce their own inference costs. We are already seeing initial steps in this direction, such as Google's Gemini 3 Flash (a smaller, faster model) outperforming Gemini 3 Pro (the larger flagship) in key benchmarks, including coding and long-horizon tasks. Notably, Google itself is heavily promoting the Flash model. For this reason, I expect significant activity in this space in 2026.

5. The Discourse Gap: Talking vs Building

Meaningful discussion of guardrails, bias, or the benefits of AI reliability is greatly enhanced by hands-on deployment experience. Yet much of the AI discourse we encounter, particularly in social media and conference circuits, can drift from the messy realities of production systems.

This gap between theoretical discussion and practical implementation isn't inherently problematic, but it becomes concerning when it shapes expectations and investment decisions. I am listing my favourite three here for completeness and correctness:

1. "**We need guardrails**" — of course we do, but safety is hard, iterative, trade-off-heavy work. Guardrails are not merely a few lines of code we can introduce and check off a list. It is handled from design through deployment, and I suspect we can never get it perfectly right, given the approximate nature of AI systems. In our published drone-based systems, safety was implemented through geofencing, altitude limits, classification-confidence thresholds, and human-verification checkpoints. Calibrating only the confidence threshold required substantial field testing. If it is too low, it yields false positives that waste resources; if it is too high, it misses targets. Together with colleagues from the Computer Science department, we also explored this in the context of runtime verification at the International Conference on Bridging the Gap between AI and Reality.
2. "**We should avoid biased data**" — Bias can indeed be harmful and result in unexpected output. However, bias is embedded in learned representations, not just data artefacts. In our 2023 research on face processing pipelines, we demonstrated that bias manifests at every stage (detection, alignment, representation, verification) across demographic groups. We cannot "scrub" bias post-training without degrading general capabilities. Fixing it requires retraining using representative data across the whole demographic spectrum, which means starting from scratch rather than applying a patch.
3. "**We need fairness in AI and align it to our values**" — just like the rest, this is something we all desire, but also know that it is not achievable. Put AI and computers aside. As human beings, we struggle to define (let alone measure) fairness. Alignment with values sounds great until we realise that we must choose which values we are referring to, and that every choice of value entails setting aside someone else's values. Technologically, fairness is often mathematically impossible to perfect. Optimising for one virtue (e.g., equal false-positive rates across groups) degrades another (e.g., overall accuracy). When hammering down on fairness, we are choosing which injustice to accept, not eliminating injustice. In media analysis applications, should a bias detection model optimise for equal false-positive rates across groups or for overall accuracy? You cannot optimise for both. Choosing one means accepting trade-offs in the other. There's no 'fair' solution, only transparent documentation of which trade-off you decided. This is one of the issues we are exploring within the DIMAS Project.

This all boils down to a lack of awareness and adequate education on how AI works in practice, which is why AI literacy cannot be addressed in the same way we address digital literacy. We are dealing with highly sophisticated systems that we cannot afford to speculate about.

My dilemma is the ratio of effort I dedicate to building with my teams and outreach. I generally limit my participation in events that include panels and keynotes. Out of respect towards the organisers and their audience, they deserve thorough preparation. If done well, you need at least one working day to prepare. The event itself also requires approximately another full day (travel time, the event, networking, and context switching). It follows that each panel or keynote eats away at least 40% of the week. The rest of the week includes meetings, lecturing and production/building. If this becomes a cycle or regular 'talking time', there is ultimately limited or no time for

building. Even worse, this results in diluted contributions that are not grounded in practice, thus becoming cheap noise that does not really add value.

The builders know that bias mitigation is not solved by wishing for better data. Fairness is often mathematically impossible to perfect, and fine-tuning for one virtue usually degrades another. Bias is embedded in the model's learned representations and cannot be selectively removed. Implementing safety is hard, iterative, trade-off-heavy work. It is easy to talk and speculate about, but it remains an engineering nightmare.

What Matters in 2026?

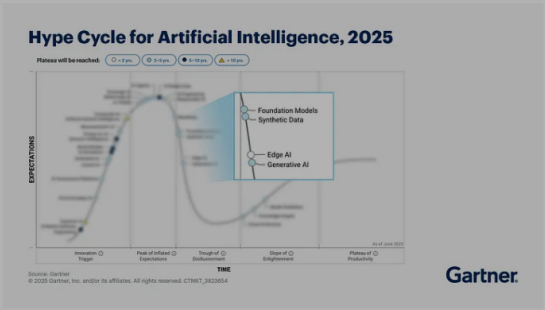
While I began this article with reports highlighting failures, the sobering takeaway from 2025 was that the industry is transitioning from magic to metrics. Industries are learning (in some cases, the hard way) that AI cannot be applied to an existing business or system. AI amplifies human capabilities; if a company knows its purpose and customers, AI will amplify it. If someone is lazy, AI amplifies this as well. We should look beneath the surface to better understand AI's effectiveness and ask: "What is AI really amplifying here?"

Below is a list of what I am personally following (and hoping for) in 2026:

1. The **per-token infrastructure cost** is critical for large-scale deployment. There is fierce competition on inference efficiency, and we all stand to benefit from its outcome.
2. Rather than horizontal wrappers, I expect **vertical adoption in high-stakes domains**. For example, rather than expecting a transformation in healthcare from AI, we will likely see AI further integrated into radiology to improve prediction and support physicians in making better judgments.
3. Organisations are realising **the value of AI Literacy and Governance**. No organisation can afford to fall behind and not have its workforce (at least) literate about AI tools and opportunities. Governance follows once there is awareness. AI should not be pushed underground. Employees are using it, and underground use is where misuse festers.
4. The gap between frontier models and "good enough" models will narrow dramatically, paving the way for **efficient, locally deployable models**. Applications currently dependent on API calls will start migrating to local inference. I expect adoption in bandwidth-constrained, privacy-sensitive, or offline-first environments where cloud dependency was previously a dealbreaker.
5. Further **decline of SEO as we know it**, as increasingly sophisticated content-analysis algorithms can focus more on what matters to humans. It will be less about the top 10 pages related to specific keywords and more about **Experience, Expertise, Authoritativeness, and Trustworthiness (EEAT)**. Startups like Polzify can help businesses further strengthen what matters to their clients.

What emerges from this correction will not be hype-worthy. This will likely feature verticalised tools that solve specific problems, governed by humans who understand both the domain and the technology's limitations.

The Shadow AI economy (individuals using ChatGPT/Gemini/Grok/Claude for email drafts and meeting summaries instead of their organisation's approved AI, if any) will persist because the value is immediate and the risk is low. But enterprise-grade, mission-critical AI deployment requires infrastructure, governance, and unit economics that most current projects lack.



Gartner's AI Hype Cycle is always a good place to start when planning your way forward.

The "AI Revolution" will more closely resemble the adoption curve of any system. It is a grinding, expensive, multi-year integration project with

system. It is a grinding, expensive, multi-year integration project with modest but real efficiency gains. Not magic. Metrics.

Conclusion

Just as with data bias, discerning the signal in AI progress amid the surrounding noise is not easy. The loudest narratives in 2026 will continue to feel like noise. These obscure the hard, incremental effort that real progress demands.

When evaluating AI commentary (online or in person), ask yourself: What has this person built (publications, products, businesses, R&D projects, etc.)?

Or is it just someone else surfing the hype wave and being reinforced by an LLM, which is convincing them that the world cannot survive without their wisdom?

. . .

Dr Dylan Seychell is a lecturer in Artificial Intelligence at the University of Malta, where he leads research projects spanning environmental AI, media transparency, and urban operations. He is a technical expert certified by the Malta Digital Innovation Authority, a co-founder of companies in the tourism sector, and the deputy chair of the Tourism Operators Business Section committee within The Malta Chamber. His work includes deploying computer vision systems with Malta's government agencies and working on the National AI Literacy Programme.

Subscribe to his newsletter for insights on AI deployment, research, and education: www.dylanseychell.com

AI · Artificial Intelligence · Business Strategy · Innovation · Business

43

Written by Dylan Seychell

253 followers · 412 following

Academic @UMmalta specialised in #AI & Computer Vision | Founder | Principal Investigator of Malta's National AI Literacy Programme 'Al Għal Kulħadd'

Follow

No responses yet

Write a response

What are your thoughts?

More from Dylan Seychell

datasets in the age of AI

evaluation done efficiently without ignoring...

Dylan Seychell · May 19

The Philosophy Behind Malta's National AI Literacy Programme

Why is it more about thinking than giving

Dylan Seychell · Jul 2, 2017

The Evolution of Users over the past Decades

When discussing Human-Computer

away AI subscriptions?

Interaction and User Experience, we...

See all from Dylan Seychell