The yardstick keeps moving

Why nobody knows exactly what an AI system can do, and why that became a business risk this year

2026

At a glance

The best documented measure of AI progress shows that the length of the tasks a model can handle has doubled roughly every three months since 2024. When the organisation behind that measure replaced its task set and its test environment in January 2026, the scores of the same models moved 57% down and 55% up. The only randomised field trial ever run with real developers found a slowdown of 19%, and the follow-up trial broke down because participants refused to work without the tools any longer. Of 445 benchmark papers examined, 16% report a margin of uncertainty beside the number. The European legislator hit the same wall and postponed the rules for high-risk AI to December 2027, because nobody had written down in time how to measure conformity.

This report is not about whether AI works. It is about whether you can establish that it does, and what it costs when you cannot.

16%

of 445 AI benchmarks examined report any measure of uncertainty beside the published number. Bean et al., NeurIPS 2025

Foreword

Every conversation about AI inside a company ends up at a number sooner or later. A percentage on a ranking, a time saving per week, a success rate on a test set. That number decides an investment, a hire, a contract clause or a tender. This report is about where those numbers come from.

It looked at the primary evaluation sources published between March 2025 and July 2026: the measurement protocols of the research organisation that maintains the best known measure of autonomous AI ability, the systematic literature review of the validity of benchmarks, the only randomised trial with professional developers, and the European legislative texts that turn measurement into a legal duty.

What those sources have in common is striking. They are written by the people who do the measuring themselves, and they are unusually candid about how shaky their own figures are. That candour rarely reaches the summary. This report puts them side by side and describes what they mean for someone who has to decide something on the basis of such a number.

Venture Campus

01 From demonstration to burden of proof

Until recently the question around an AI system was whether it worked. Someone showed something, the audience was impressed or it was not, and the matter was settled. That period is over, and the date is known. Regulation (EU) 2024/1689, the AI Act, turns an AI system into a regulated product with a conformity assessment, technical documentation and an obligation to demonstrate what it does.

From 2 August 2026 the transparency obligations of article 50 apply across the Union. Anyone communicating with an AI system must be able to know it, and anyone generating content must mark it as such. That date has not moved. For legacy systems already on the market at that moment, article 50(2) applies from 2 December 2026. The new prohibitions apply from the same date.

The consequence reaches further than the legislation. A tender asks for performance figures. An insurer asks for a margin of error. A customer asks how often it goes wrong. A judge, as the treatment of the regulation further on shows, asks for a standard to test against. All of those questions come down to the same thing: prove it.

And that is where the problem starts, because the apparatus of proof has not grown along with the technology. The figures the sector describes itself with come largely from test set-ups that nobody built with the thought that a contract, a fine or a court case might one day hang on them. This report is about that apparatus: what it measures, what it does not measure, and what happens when you rely on it.

Source: Regulation (EU) 2024/1689 (AI Act), article 50 · Council of the European Union, press release on the final approval of the digital omnibus on AI, 29 June 2026

02 The best yardstick there is

The most serious attempt to measure AI ability over time comes from METR, an American non-profit research foundation. Its measure is called the time horizon, and the idea is elegant: do not measure how many questions a model answers correctly, but how long a task may take before the model fails. For every task in the set, the time a human professional spends on it is measured. The horizon of a model is the task length at which it succeeds half of the time.

That measure grows exponentially, and it has done so for six years. In the original publication of March 2025 the horizon of the best model doubled roughly every 196 days, a little over seven months, across the period 2019 to 2025. After the revision of January 2026 those seven months hold for the whole series, but the sub-periods draw a steeper picture: 130.8 days since 2023, and 88.6 days since 2024. That last figure is less than three months.

> The best known measure of AI progress records how long a task may take before the model fails half of the time.

Read that figure precisely, because the precision sits in the word that easily falls away. The horizon is defined at a success rate of fifty per cent. It is the task length at which the model fails as often as it succeeds. On a task of four minutes the current generation reaches almost one hundred per cent; on tasks that cost a human more than four hours, the success rate drops below ten per cent. What a company needs almost never sits at fifty per cent. Bookkeeping needs 99.9%. So does a medical report.

METR is exceptionally clear about that limitation. The tasks are designed to stand on their own and to be fully specified, so that they are fair to human and machine alike. Real work is not like that: it leans on earlier conversations, on tacit knowledge, on familiarity with an existing code base. A two-hour task in the set corresponds, according to METR, more closely to what a new employee or a hired freelancer finishes in two hours than to what an experienced colleague does in the same time. And the set consists predominantly of programming, machine learning and security tasks.

Source: METR, Measuring AI Ability to Complete Long Tasks, 19 March 2025 · METR, Time Horizon 1.1, 29 January 2026 · Kwa et al., arXiv:2503.14499

03 What happened when the meter itself changed

On 29 January 2026 METR published a new version of its estimates. Two things had changed. The task set went from 170 to 228 tasks: 73 added, 15 removed, 53 adjusted because the description was confusing, the scoring function contained errors or the task was too easy to exploit. And the test environment moved from Vivaria, which METR had built itself, to Inspect, the open-source framework of the British AI Security Institute.

The models had not changed. The figures had. The estimated horizon of GPT-4 from November 2023 fell by 57%. That of an older GPT-4 fell by 35%. GPT-5 rose by 55%. Claude Opus 4.5 rose by 11%. The move of test environment alone made the horizon of GPT-4o fall from 9.2 to 6.0 minutes, and for two models the difference between the two environments was statistically significant: the same question, different scaffolding, demonstrably different result.

One more detail from the same publication, and it is the most important one in the whole report. The set contains 31 tasks estimated to take eight hours or more. Of those 31, a human baseline time has actually been measured for five. The other 26 are estimates. The trend line the sector cites as evidence of an approaching upheaval rests at the top end, precisely where it becomes interesting, on estimated rather than measured human time.

None of this is an indictment of METR. The opposite: no supplier publishes this kind of self-criticism, and it is known only because METR wrote it down itself. The lesson is methodological. The measured ability of a model is not a property of the model. It is a property of the model, the scaffolding around it, the task set, the scoring function and the human reference time together. Change one of them, and the figure changes by tens of per cent without anything at all having happened to the model.

5 of 31

of METR's longest tasks (eight hours or more) has a measured human baseline time. The other 26 are estimates. Time Horizon 1.1, 29 January 2026

Source: METR, Time Horizon 1.1, 29 January 2026, including the comparison tables TH1 against TH1.1 and Vivaria against Inspect

04 What the literature on benchmarks says

If the most careful measurement in the field is this sensitive to its own set-up, what does that say about the rest? Since late 2025 there is an answer to that question. A team of 29 peer reviewers, with researchers from Oxford, Yale and Stanford among others, screened 46,114 papers from six major conferences in machine learning and language technology, kept 445 that introduce a benchmark, and assessed those one by one on construct validity: does this thing measure what it claims to measure?

Almost every paper shows a weakness on at least one point. Sixteen per cent uses an uncertainty estimate or a statistical test to compare results. Twenty-seven per cent uses a sample assembled on the basis of availability. Core notions such as safety and robustness are often not defined or not operationalised, which makes the conclusions about them untenable. The work appeared at NeurIPS 2025 and ends with eight recommendations and a checklist, which in itself says how basic the problem is.

> Sixteen in a hundred benchmark papers put a margin of uncertainty beside their own number.

Beneath that lies a second layer: the benchmarks themselves contain errors. A re-annotation of MMLU, one of the most cited test sets in the world, found that 6.49% of the questions contain an error. An audit of SWE-bench, the standard set for programming tasks, reported 32.7% solution leakage: the solution was findable in the problem statement in one way or another. The Joint Research Centre of the European Commission arrived in 2025 at nine systemic problems in AI benchmarking, among them misplaced incentives and the possibility of gaming the test.

The consequence is not that benchmarks are worthless. The consequence is that a difference of a few percentage points on a ranking has no meaning you can use, because in most cases no margin of error accompanies it and nobody has demonstrated that the test measures what the name promises. Two models with an identical top score can have radically different strengths and weaknesses, and an average hides that by definition.

Source: Bean et al., Measuring what Matters: Construct Validity in Large Language Model Benchmarks, NeurIPS 2025 Datasets and Benchmarks Track, arXiv:2511.04703 · Gema et al., re-annotation of MMLU · Aleithan et al., audit of SWE-bench · Eriksson et al., meta-review commissioned by the Joint Research Centre of the European Commission, 2025

05 The field trial that broke down

Benchmarks measure what a model can do in a laboratory. The question a company asks is a different one: does my team get faster? On that there is exactly one randomised trial in real conditions. METR recruited sixteen experienced developers in 2025 from mature open-source projects, who had worked on their own code base for five years on average. They supplied the

tasks themselves from their own backlog: 246 of them, each assigned at random to AI or no AI. Cursor Pro and Claude 3.5 and 3.7 Sonnet were allowed, the frontier at the time.

Beforehand the participants predicted that AI would make them 24% faster. Afterwards, once it was over, they estimated that it had made them 20% faster. Measured, they were 19% slower. The confidence interval ran from 2% to 39% slowdown. The researchers wrote plainly that they had expected a speed-up. The most striking thing is not the slowdown but the gap: people who had undergone the slowdown themselves still believed afterwards the opposite of what the clock had recorded.

Then came the second round, and it is more instructive than the first. METR started again in August 2025, with 57 developers across 143 code bases and more than 800 tasks. On 24 February 2026 the organisation published that it considered its own result a poor approximation and would change the trial design. The reason: the control group no longer exists. A growing share of the developers do not want to take part because they do not want to do half of their work without AI. Between 30% and 50% indicated that they deliberately did not submit certain tasks, out of fear that these would be assigned to the no-AI arm. One participant completed no task at all in that arm.

The raw figures meanwhile point the other way. For the participants from the first study METR now estimates a speed-up of 18%, with an interval from 38% speed-up to 9% slowdown. For the newly recruited group that is 4%, with an interval from 15% speed-up to 9% slowdown. Both intervals cut through zero. And that is the heart of it: the effect has probably become positive, and it is no longer measurable with the method that once measured it as negative, because people no longer let themselves be randomised. One participant described working without the tools as walking through the city after getting used to the taxi.

Source: Becker, Rush, Barnes and Rein, Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity, METR, 10 July 2025, arXiv:2507.09089 · Becker, Rush, Cunningham, Rein and Mahamud, We are Changing our Developer Productivity Experiment Design, METR, 24 February 2026

06 The legislator hits the same wall

The AI Act rests on a European construction that has worked for decades in machinery and toys: the law formulates essential requirements, and harmonised standards translate those into testable specifications. Whoever follows the standard enjoys a presumption of conformity. For AI, CEN-CENELEC set up technical committee JTC 21 for that purpose, founded on 1 June 2021, by now more than 300 experts from more than twenty countries, with standardisation request M/593 and amendment M/613 as its mandate.

Those standards have not arrived. The original deadline of 30 April 2025 was formally moved by the Commission to 31 August 2025 and that one was missed too. prEN 18286, the standard for the quality management system and the hinge the others hang on, only went into public enquiry on 30 October 2025, eight months late. In October 2025 the technical boards of CEN and CENELEC reached for an exceptional procedure: on a positive enquiry vote publication may follow directly, without a separate formal vote, and a small editing group would finish six of the slowest drafts. Target: available by the fourth quarter of 2026.

The reason it takes so long is precisely the subject of this report. Part of what the AI Act asks for does not yet exist as state of the art. The regulator asks for standardisation in areas where there is no standard to standardise, and a consensus process cannot organise consensus about a measurement nobody can carry out. Without a standard, a legal vacuum would have opened from August 2026: obligations without a way to know whether you meet them.

On 19 November 2025 the Commission made a proposal that tied the application of the high-risk rules to the availability of supporting means, with fixed end dates as a backstop. Parliament and Council reached a provisional agreement on 7 May 2026, Parliament approved it on 16 June 2026 and the Council gave the final green light on 29 June 2026. The outcome is set out below. It is worth looking at, because it is a legislator establishing that it cannot regulate what it cannot yet measure, and taking sixteen months for that.

Date	What starts to apply
2 August 2026	Transparency obligations of article 50 (unchanged)
2 December 2026	Article 50(2) for existing systems; new prohibited practices
2 August 2027	At least one national AI regulatory sandbox per member state
2 December 2027	High-risk obligations for standalone systems (annex III), previously 2 August 2026
2 August 2028	High-risk obligations for AI embedded in regulated products (annex I)

Application dates of the AI Act after the digital omnibus, approved on 29 June 2026

Source: European Commission, proposal for a digital omnibus on AI, 19 November 2025 · Council of the EU and European Parliament, provisional agreement, 7 May 2026; Parliament 16 June 2026; Council, final approval, 29 June 2026 · CEN-CENELEC, Update on CEN and CENELEC's Decision to Accelerate the Development of Standards for Artificial Intelligence, 23 October 2025 · CEN-CLC/JTC 21, standardisation request M/593 and amendment M/613

07 What it changes for companies

The first inference is unpleasant and unavoidable: your evidence is not for sale. A score on a public ranking is a statement about the task set, the scaffolding and the scoring function of whoever ran the test. What that figure predicts about your tasks, your data and your reliability threshold, nobody has established. The only evaluation that says anything about your application is one on your own work. That is no longer a side activity but a cost with a budget line of its own.

The second is that the threshold determines the investment, not the model. The difference between a horizon at fifty per cent success and one you can actually use is the whole project. For a task where an error is cheap and visible, fifty per cent is usable. For a task where an error is expensive and silent, ninety-nine per cent is the minimum, and the road there does not run via a bigger model but via control, verification and the design of the process around it. That work is the real innovation, and it is precisely the work a demonstration does not show.

> A score on a public ranking is a statement about the supplier's test set-up, not about your work.

The third is that self-reporting is not evidence. The METR trial delivered the sharpest illustration of that the literature has: people who were 19% slower believed afterwards that they had been

20% faster. Anyone who justifies an AI investment with a satisfaction survey among the users measures the feeling, not the effect. That feeling went wrong systematically and by a wide margin in the wrong direction, among experts, in a controlled set-up, with the clock running.

The fourth is a diary appointment. The sixteen months of postponement for the high-risk obligations are not a pause, they are a moved deadline, and the transparency obligations of 2 August 2026 have not moved. Anyone who builds a system between now and December 2027 that falls under annex III is building against standards that are still in draft at this moment and that are scheduled to become available only at the end of 2026. The design of your evaluation is therefore itself a bet on the final text of a standard.

Source: METR, arXiv:2507.09089, 10 July 2025 · Regulation (EU) 2024/1689, article 40 (presumption of conformity) and article 50 · METR, Time Horizon 1.1, 29 January 2026, on the 50% and 80% horizon

08 What does not add up yet

The most honest thing to be said about this subject is that the direction of the productivity effect is still not established with certainty. The only randomised trial found a slowdown, and its own authors now consider that figure outdated. The follow-up trial cannot be interpreted. What remains is self-reporting, and that is precisely the instrument that went badly wrong in the same study. METR's survey of 11 May 2026 among 349 technical workers finds a median self-reported value increase of 1.4 to 2 times, and adds itself that there are reasons to doubt that order of magnitude.

The trend line deserves the same reticence. The acceleration to a doubling time of 88.6 days is measured over one year. The revised series contains fourteen models where the previous one counted thirty-three. The confidence intervals have become narrower but remain wide: the upper bound for the best model still sits at well over double the point estimate. METR writes itself that the number of tasks the newest generation cannot handle has become small, and that the yardstick therefore threatens to saturate at the top end. An extrapolation on such a basis is a hypothesis, not a forecast.

The observational figures put forward as an alternative also measure something other than what they appear to measure. That roughly 4% of the commits on GitHub are said to be written by an agent is an adoption figure. It says nothing about the quality, the amount of rework or the net time saved. Some participants in the METR trial noted that they took on different tasks with agents than without, and that the quality of the end result differed between the two arms. If the task changes, a time comparison measures nothing any more.

Finally, a tension the sector has not talked out. The high-risk rules were postponed because the standards were missing. From December 2027 those same standards are the measurement, with a presumption of conformity as the legal effect. They are now being finished under an exceptional procedure that skips the formal vote, and within JTC 21 objection has been raised against that openly, with a request to revise the measure. A standard that comes about via a shortened route and then decides whether a product may legally be on the market is an invitation to litigate. Anyone who builds their conformity on those standards in 2027 builds on a foundation that is still under discussion.

Source: METR, Measuring the Self-Reported Impact of Early-2026 AI on Technical Worker Productivity, 11 May 2026 · METR, Time Horizon 1.1, 29 January 2026 · METR, We are Changing our Developer Productivity Experiment Design, 24 February 2026 · Cantero Gamito, From Consensus to Exceptionality: What the EU's AI Standards Crisis Reveals About Delegated Technical Governance, 28 November 2025

About Venture Campus

Venture Campus is a Belgian consultancy for innovation funding. The firm prepares applications, writes them and submits them together with the client to the Flemish, federal and European instruments.

The assessment criteria are sector-neutral. All sectors, products and services qualify. The client brings the technology, the domain and the plan. Venture Campus brings the scheme, its criteria and the evaluator's perspective, applied before submission.

For the work this report describes, building knowledge and removing technical uncertainty, the O&O instrument of VLAIO is usually the appropriate channel: 25 to 45% of the accepted costs for development projects, 25 to 60% for research projects, with a maximum of 3 million euros. VLAIO funds knowledge building and technical risk, not products. The percentages apply on the date on the cover.

On most assignments no cure, no pay applies: you pay when the support has been granted. That does not hold for Strategische Transformatiesteun and it does not describe the fiscal trajectories. The analysis beforehand is free, as is the subsidiescan with six questions. If there is no realistic case, the firm says so.

Venture Campus BV | Schaarbeekstraat 20E bus 11 | 9120 Beveren, Belgium | +32 333 11 333 | venturecampus.com

SOURCES

Regulation (EU) 2024/1689 (AI Act), articles 40 and 50 · Council of the European Union, press release on the final approval of the digital omnibus on AI, 29 June 2026 · Council of the EU and European Parliament, provisional agreement on the digital omnibus on AI, 7 May 2026; approval European Parliament 16 June 2026 · European Commission, proposal for a regulation simplifying the implementation of harmonised rules on AI (digital omnibus on AI), 19 November 2025 · CEN-CENELEC, Update on CEN and CENELEC's Decision to Accelerate the Development of Standards for Artificial Intelligence, 23 October 2025 · CEN-CLC/JTC 21, standardisation request M/593 and amendment M/613; draft standard prEN 18286, public enquiry from 30 October 2025 · METR, Measuring AI Ability to Complete Long Tasks, 19 March 2025 (Kwa et al., arXiv:2503.14499) · METR, Time Horizon 1.1, 29 January 2026 · METR, Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity, 10 July 2025 (Becker, Rush, Barnes and Rein, arXiv:2507.09089) · METR, We are Changing our Developer Productivity Experiment Design, 24 February 2026 · METR, Measuring the Self-Reported Impact of Early-2026 AI on Technical Worker Productivity, 11 May 2026 · Bean et al., Measuring what Matters: Construct Validity in Large Language Model Benchmarks, NeurIPS 2025 Datasets and Benchmarks Track, arXiv:2511.04703 · Gema et al., re-annotation of MMLU; Aleithan et al., audit of SWE-bench · Eriksson et al., meta-review of AI benchmarking, Joint Research Centre of the European Commission, 2025 · Cantero Gamito, From Consensus to Exceptionality: What the EU's AI Standards Crisis Reveals About Delegated Technical Governance, 28 November 2025