L’étalon bouge aussi

Pourquoi personne ne sait précisément ce qu’un système d’IA sait faire, et pourquoi c’est devenu un risque d’entreprise cette année

2026

En un coup d’oeil

La mesure la mieux documentée du progrès de l’IA montre que la durée des tâches qu’un modèle sait mener à bien double environ tous les trois mois depuis 2024. Lorsque l’organisation qui tient cette mesure a remplacé sa série de tâches et son environnement de test en janvier 2026, les scores des mêmes modèles se sont déplacés de 57% vers le bas et de 55% vers le haut. Le seul essai de terrain randomisé jamais mené avec de vrais développeurs a constaté un ralentissement de 19%, et l’essai de suivi s’est enrayé parce que les participants refusaient encore de travailler sans l’outil. Sur 445 articles de benchmark examinés, 16% indiquent une marge d’incertitude à côté du chiffre. Le législateur européen a buté sur le même mur et a reporté à décembre 2027 les règles pour l’IA à haut risque, parce que personne n’avait écrit à temps comment se mesure la conformité.

Ce rapport ne traite pas de la question de savoir si l’IA fonctionne. Il traite de la question de savoir si vous pouvez l’établir, et de ce que cela coûte lorsque vous n’y arrivez pas.

16%

des 445 benchmarks d’IA examinés indiquent une mesure d’incertitude quelconque à côté du chiffre publié. Bean et al., NeurIPS 2025

Avant-propos

Toute conversation sur l’IA dans une entreprise finit tôt ou tard sur un chiffre. Un pourcentage dans un classement, un gain de temps par semaine, un taux de réussite sur une série de tests. Ce chiffre décide d’un investissement, d’un recrutement, d’une clause contractuelle ou d’un marché public. Ce rapport porte sur l’origine de ces chiffres.

Il examine les sources d’évaluation primaires publiées entre mars 2025 et juillet 2026: les protocoles de mesure de l’organisation de recherche qui tient la mesure la plus connue de la capacité autonome de l’IA, la revue systématique de la littérature sur la validité des benchmarks, le seul essai randomisé mené avec des développeurs professionnels, et les textes législatifs européens qui font de la mesure une obligation juridique.

Ce que ces sources ont en commun est frappant. Elles sont écrites par les personnes qui mesurent elles-mêmes, et elles sont d’une franchise inhabituelle sur la fragilité de leurs propres chiffres. Cette franchise atteint rarement le résumé. Ce rapport les rassemble et décrit ce qu’elles signifient pour quelqu’un qui doit décider sur la base d’un tel chiffre.

Venture Campus

01 De la démonstration à la charge de la preuve

Jusqu’il y a peu, la question autour d’un système d’IA était de savoir s’il fonctionnait. Quelqu’un montrait quelque chose, le public était impressionné ou ne l’était pas, et l’affaire était réglée. Cette période est terminée, et la date est connue. Le règlement (UE) 2024/1689, le règlement sur l’IA, fait d’un système d’IA un produit réglementé, avec une évaluation de la conformité, une documentation technique et une obligation de démontrer ce qu’il fait.

À partir du 2 août 2026, les obligations de transparence de l’article 50 s’appliquent dans toute l’Union. Qui communique avec un système d’IA doit pouvoir le savoir, et qui génère du contenu doit le marquer comme tel. Cette date n’a pas bougé. Pour les systèmes existants qui sont déjà sur le marché à ce moment, l’article 50, paragraphe 2 s’applique à partir du 2 décembre 2026. Les nouvelles interdictions s’appliquent à partir de la même date.

La conséquence dépasse la législation. Un marché public demande des chiffres de performance. Un assureur demande une marge d’erreur. Un client demande à quelle fréquence cela se passe mal. Un juge, comme le montre plus loin le traitement de la réglementation, demande une norme à laquelle il peut confronter le produit. Toutes ces questions reviennent au même: prouvez-le.

Et c’est là que le problème commence, car l’appareil de preuve n’a pas grandi avec la technologie. Les chiffres par lesquels le secteur se décrit proviennent en grande partie de dispositifs de test que personne n’a construits en pensant qu’un contrat, une amende ou un procès y serait un jour suspendu. Ce rapport porte sur cet appareil: ce qu’il mesure, ce qu’il ne mesure pas, et ce qui arrive lorsque vous vous y appuyez.

Source : Règlement (UE) 2024/1689 (règlement sur l’IA), article 50 · Conseil de l’Union européenne, communiqué sur l’approbation définitive de l’omnibus numérique sur l’IA, 29 juin 2026

02 Le meilleur étalon qui existe

La tentative la plus sérieuse de mesurer la capacité de l’IA dans le temps vient de METR, une fondation de recherche américaine sans but lucratif. Sa mesure s’appelle l’horizon temporel, et l’idée est élégante: ne mesurez pas combien de questions un modèle résout correctement, mais combien de temps une tâche peut durer avant que le modèle n’échoue. Pour chaque tâche de la série, on mesure le temps qu’un professionnel humain y consacre. L’horizon d’un modèle est la longueur de tâche à laquelle il réussit une fois sur deux.

Cette mesure croît de façon exponentielle, et elle le fait depuis six ans. Dans la publication d’origine de mars 2025, l’horizon du meilleur modèle doublait environ tous les 196 jours, un peu plus de sept mois, sur la période 2019 à 2025. Après la révision de janvier 2026, ces sept mois tiennent pour toute la série, mais les sous-périodes dessinent une image plus raide: 130,8 jours depuis 2023, et 88,6 jours depuis 2024. Ce dernier chiffre fait moins de trois mois.

> La mesure la plus connue du progrès de l’IA enregistre combien de temps une tâche peut durer avant que le modèle n’échoue une fois sur deux.

Lisez ce chiffre avec précision, car la précision se loge dans le mot qui tombe facilement. L’horizon est défini à un taux de réussite de cinquante pour cent. C’est la longueur de tâche à laquelle le modèle échoue aussi souvent qu’il réussit. Sur une tâche de quatre minutes, la génération actuelle atteint presque cent pour cent; sur des tâches qui coûtent à un humain plus de quatre heures, le taux de réussite tombe sous les dix pour cent. Ce dont une entreprise a besoin ne se situe presque jamais à cinquante pour cent. La comptabilité a besoin de 99,9%. Un rapport médical aussi.

METR est particulièrement clair sur cette limite. Les tâches sont conçues pour tenir debout seules et pour être entièrement spécifiées, afin d’être équitables pour l’humain comme pour la machine. Le travail réel n’est pas ainsi: il s’appuie sur des conversations antérieures, sur des connaissances tacites, sur la familiarité avec une base de code existante. Une tâche de deux heures dans la série correspond, selon METR, davantage à ce qu’un nouveau collaborateur ou un indépendant engagé boucle en deux heures qu’à ce qu’un collègue expérimenté fait dans le même temps. Et la série est composée en majorité de travail de programmation, d’apprentissage automatique et de tâches de sécurité.

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

03 Ce qui est arrivé quand le compteur lui-même a changé

Le 29 janvier 2026, METR a publié une nouvelle version de ses estimations. Deux choses avaient changé. La série de tâches est passée de 170 à 228 tâches: 73 ajoutées, 15 supprimées, 53 adaptées parce que l’énoncé prêtait à confusion, que la fonction de score contenait des erreurs ou que la tâche était trop facile à exploiter. Et l’environnement de test a déménagé de Vivaria, que METR avait construit lui-même, vers Inspect, le cadre open source de l’AI Security Institute britannique.

Les modèles n’avaient pas changé. Les chiffres, si. L’horizon estimé de GPT-4 de novembre 2023 a chuté de 57%. Celui d’un GPT-4 plus ancien a chuté de 35%. GPT-5 a augmenté de 55%. Claude Opus 4.5 a augmenté de 11%. Le seul déménagement d’environnement de test a fait tomber l’horizon de GPT-4o de 9,2 à 6,0 minutes, et pour deux modèles la différence entre les deux environnements était statistiquement significative: la même question, un autre pilotage, un résultat démontrablement différent.

Encore un détail de la même publication, et c’est le plus important de tout le rapport. La série contient 31 tâches estimées à huit heures ou plus. Sur ces 31, un temps de référence humain a réellement été mesuré pour cinq. Les 26 autres sont des estimations. La courbe de tendance que le secteur cite comme preuve d’un bouleversement imminent s’appuie, dans le haut, précisément là où cela devient intéressant, sur du temps humain estimé et non mesuré.

Rien de tout cela n’est un réquisitoire contre METR. Au contraire: aucun fournisseur ne publie ce genre d’autocritique, et on ne le sait que parce que METR l’a écrit lui-même. La leçon est méthodologique. La capacité mesurée d’un modèle n’est pas une propriété du modèle. C’est une

propriété du modèle, du pilotage autour de lui, de la série de tâches, de la fonction de score et du temps de référence humain ensemble. Changez-en un, et le chiffre change de dizaines de pour cent, sans que quoi que ce soit ne soit arrivé au modèle.

5 sur 31

des tâches les plus longues de METR (huit heures ou plus) ont un temps de référence humain mesuré. Les 26 autres sont des estimations. Time Horizon 1.1, 29 janvier 2026

Source : METR, Time Horizon 1.1, 29 janvier 2026, y compris les tableaux de comparaison TH1 contre TH1.1 et Vivaria contre Inspect

04 Ce que dit la littérature sur les benchmarks

Si la mesure la plus soigneuse du domaine est à ce point sensible à son propre dispositif, que dire du reste? À cette question il existe une réponse depuis fin 2025. Une équipe de 29 relecteurs spécialisés, avec des chercheurs d’Oxford, de Yale et de Stanford entre autres, a passé au crible 46 114 articles de six grandes conférences en apprentissage automatique et en traitement du langage, en a retenu 445 qui introduisent un benchmark, et les a évalués un par un sur la validité de construit: cette chose mesure-t-elle ce qu’elle prétend mesurer?

Presque chaque article présente une faiblesse sur au moins un point. Seize pour cent utilisent une estimation d’incertitude ou un test statistique pour comparer les résultats. Vingt-sept pour cent utilisent un échantillon constitué sur la base de la disponibilité. Des notions centrales comme la sécurité et la robustesse ne sont souvent ni définies ni opérationnalisées, ce qui rend intenables les conclusions à leur sujet. Le travail est paru à NeurIPS 2025 et se termine par huit recommandations et une liste de contrôle, ce qui dit en soi à quel point le problème est élémentaire.

> Seize articles de benchmark sur cent placent une marge d’incertitude à côté de leur propre chiffre.

En dessous se trouve une deuxième couche: les benchmarks eux-mêmes contiennent des erreurs. Une réannotation de MMLU, l’une des séries de tests les plus citées au monde, a trouvé que 6,49% des questions contiennent une erreur. Un audit de SWE-bench, la série standard pour les tâches de programmation, a signalé 32,7% de fuite de solution: la solution était, d’une manière ou d’une autre, trouvable dans l’énoncé. Le Centre commun de recherche de la Commission européenne est arrivé en 2025 à neuf problèmes systémiques dans le benchmarking de l’IA, dont des incitants mal placés et la possibilité de jouer le test.

La conséquence n’est pas que les benchmarks sont sans valeur. La conséquence est qu’un écart de quelques points de pourcentage dans un classement n’a pas de signification que vous pouvez utiliser, parce que dans la plupart des cas aucune marge d’erreur ne l’accompagne et que personne n’a démontré que le test mesure ce que le nom promet. Deux modèles ayant un score identique en tête peuvent avoir des forces et des faiblesses radicalement différentes, et une moyenne le cache par définition.

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., réannotation de MMLU · Aleithan et al., audit de SWE-bench · Eriksson et al., méta-revue commandée par le Centre commun de recherche de la Commission européenne, 2025

05 L’essai de terrain qui s’est enrayé

Les benchmarks mesurent ce qu’un modèle sait faire dans un laboratoire. La question que pose une entreprise est autre: mon équipe en devient-elle plus rapide? Sur ce point il existe exactement un essai randomisé en conditions réelles. METR a recruté en 2025 seize développeurs expérimentés de projets open source matures, qui travaillaient en moyenne depuis cinq ans sur leur propre base de code. Ils ont fourni eux-mêmes les tâches depuis leur propre arriéré: 246 au total, chacune attribuée au hasard avec ou sans IA. Étaient autorisés Cursor Pro et Claude 3.5 et 3.7 Sonnet, l’avant-garde de l’époque.

Avant l’essai, les participants prévoyaient que l’IA les rendrait 24% plus rapides. Après coup, une fois terminé, ils estimaient qu’elle les avait rendus 20% plus rapides. Mesurés, ils étaient 19% plus lents. L’intervalle de confiance allait de 2% à 39% de ralentissement. Les chercheurs ont écrit sans détour qu’ils attendaient une accélération. Le plus frappant n’est pas le ralentissement mais l’écart: des gens qui avaient subi eux-mêmes le ralentissement croyaient encore, après coup, le contraire de ce que la montre avait enregistré.

Vint ensuite le deuxième tour, et il est plus instructif que le premier. METR a recommencé en août 2025, avec 57 développeurs sur 143 bases de code et plus de 800 tâches. Le 24 février 2026, l’organisation a publié qu’elle considérait son propre résultat comme une mauvaise approximation et qu’elle modifierait le design de l’essai. La raison: le groupe de contrôle n’existe plus. Une part croissante des développeurs ne veut pas participer parce qu’ils ne veulent pas faire la moitié de leur travail sans IA. Entre 30% et 50% ont indiqué qu’ils ne soumettaient sciemment pas certaines tâches, de peur qu’elles ne soient attribuées au bras sans IA. Un participant n’a achevé aucune tâche dans ce bras.

Les chiffres bruts pointent entre-temps dans l’autre sens. Pour les participants de la première étude, METR estime maintenant une accélération de 18%, avec un intervalle allant de 38% d’accélération à 9% de ralentissement. Pour le groupe nouvellement recruté, c’est 4%, avec un intervalle allant de 15% d’accélération à 9% de ralentissement. Les deux intervalles traversent zéro. Et c’est là le coeur: l’effet est vraisemblablement devenu positif, et il n’est plus mesurable avec la méthode qui l’a jadis mesuré comme négatif, parce que les gens ne se laissent plus randomiser. Un participant a décrit le travail sans l’outil comme le fait de traverser la ville à pied après s’être habitué au taxi.

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

06 Le législateur bute sur le même mur

Le règlement sur l’IA repose sur une construction européenne qui fonctionne depuis des décennies pour les machines et les jouets: la loi formule des exigences essentielles, et des normes harmonisées les traduisent en spécifications vérifiables. Qui suit la norme bénéficie

d'une présomption de conformité. Pour l'IA, le CEN-CENELEC a reçu à cette fin le comité technique JTC 21, créé le 1er juin 2021, aujourd'hui plus de 300 experts de plus de vingt pays, avec la demande de normalisation M/593 et l'amendement M/613 comme mandat.

Ces normes ne sont pas arrivées. L'échéance initiale du 30 avril 2025 a été formellement déplacée par la Commission au 31 août 2025, et celle-là aussi a été manquée. prEN 18286, la norme pour le système de gestion de la qualité et la charnière à laquelle les autres sont accrochées, n'est entrée en enquête publique que le 30 octobre 2025, avec huit mois de retard. En octobre 2025, les bureaux techniques du CEN et du CENELEC ont eu recours à une procédure d'exception: en cas de vote d'enquête positif, la publication peut suivre directement, sans vote formel distinct, et un petit groupe de rédaction achèverait six des projets les plus lents. Objectif: disponibles pour le quatrième trimestre de 2026.

La raison de cette lenteur est précisément le sujet de ce rapport. Une partie de ce que demande le règlement sur l'IA n'existe pas encore comme état de la technique. Le régulateur demande une normalisation dans des domaines où il n'y a pas de norme à normaliser, et un processus de consensus ne peut pas organiser un consensus autour d'une mesure que personne ne sait exécuter. Sans norme, un vide juridique se serait ouvert à partir d'août 2026: des obligations sans moyen de savoir si vous les respectez.

Le 19 novembre 2025, la Commission a fait une proposition qui liait l'application des règles à haut risque à la disponibilité des moyens de soutien, avec des dates de fin fixes comme filet. Le Parlement et le Conseil ont conclu un accord provisoire le 7 mai 2026, le Parlement l'a approuvé le 16 juin 2026 et le Conseil a donné le feu vert définitif le 29 juin 2026. Le résultat figure ci-dessous. Il vaut le coup d'oeil, car c'est un législateur qui constate qu'il ne peut pas réguler ce qu'il ne sait pas encore mesurer, et qui prend seize mois pour cela.

Date	Ce qui commence à s'appliquer
2 août 2026	Obligations de transparence de l'article 50 (inchangées)
2 décembre 2026	Article 50, paragraphe 2 pour les systèmes existants; nouvelles pratiques interdites
2 août 2027	Au moins un bac à sable réglementaire national pour l'IA par État membre
2 décembre 2027	Obligations à haut risque pour les systèmes autonomes (annexe III), auparavant 2 août 2026
2 août 2028	Obligations à haut risque pour l'IA intégrée dans des produits réglementés (annexe I)

Dates d'application du règlement sur l'IA après l'omnibus numérique, approuvé le 29 juin 2026

Source : Commission européenne, proposition d'omnibus numérique sur l'IA, 19 novembre 2025 · Conseil de l'UE et Parlement européen, accord provisoire, 7 mai 2026; Parlement 16 juin 2026; Conseil, approbation définitive, 29 juin 2026 · CEN-CENELEC, Update on CEN and CENELEC's Decision to Accelerate the Development of Standards for Artificial Intelligence, 23 octobre 2025 · CEN-CLC/JTC 21, demande de normalisation M/593 et amendement M/613

07 Ce que cela change pour les entreprises

La première conclusion est désagréable et inévitable: votre preuve n'est pas à vendre. Un score dans un classement public est une affirmation sur la série de tâches, le pilotage et la fonction de score de celui qui a fait passer le test. Ce que ce chiffre prédit de vos tâches, de vos données et de votre seuil de fiabilité, personne ne l'a établi. La seule évaluation qui dise quelque chose de

votre application est une évaluation sur votre propre travail. Ce n'est plus une activité annexe mais un poste de coût avec sa propre ligne budgétaire.

La deuxième est que le seuil détermine l'investissement, pas le modèle. La différence entre un horizon à cinquante pour cent de réussite et un horizon que vous pouvez réellement utiliser, c'est tout le projet. Pour une tâche où une erreur est bon marché et visible, cinquante pour cent est utilisable. Pour une tâche où une erreur est coûteuse et silencieuse, quatre-vingt-dix-neuf pour cent est le minimum, et le chemin pour y arriver ne passe pas par un modèle plus grand mais par le contrôle, la vérification et la conception du processus autour. Ce travail est l'innovation véritable, et c'est précisément le travail qu'une démonstration ne montre pas.

> Un score dans un classement public est une affirmation sur le dispositif de test du fournisseur, pas sur votre travail.

La troisième est que l'autodéclaration n'est pas une preuve. L'essai de METR en a livré l'illustration la plus nette que connaisse la littérature: des gens qui étaient 19% plus lents croyaient après coup qu'ils avaient été 20% plus rapides. Qui justifie son investissement en IA par une enquête de satisfaction auprès des utilisateurs mesure le sentiment, pas l'effet. Ce sentiment s'est trompé de manière systématique et ample dans la mauvaise direction, chez des experts, dans un dispositif contrôlé, montre en main.

La quatrième est un rendez-vous d'agenda. Les seize mois de report des obligations à haut risque ne sont pas une pause, ce sont une échéance déplacée, et les obligations de transparence du 2 août 2026 n'ont pas été déplacées. Qui construit d'ici à décembre 2027 un système qui relève de l'annexe III construit contre des normes qui sont pour l'instant encore à l'état de projet et qui, selon le calendrier, ne seront disponibles que fin 2026. La conception de votre évaluation est donc elle-même un pari sur le texte définitif d'une norme.

Source : METR, arXiv:2507.09089, 10 juillet 2025 · Règlement (UE) 2024/1689, article 40 (présomption de conformité) et article 50 · METR, Time Horizon 1.1, 29 janvier 2026, sur l'horizon à 50% et à 80%

08 Ce qui ne colle pas encore

La chose la plus honnête à dire sur ce sujet est que le sens de l'effet de productivité n'est toujours pas établi avec certitude. Le seul essai randomisé a trouvé un ralentissement, et ses propres auteurs considèrent désormais ce chiffre comme dépassé. L'essai de suivi n'est pas interprétable. Ce qui reste est l'autodéclaration, et c'est précisément l'instrument qui a échoué lamentablement dans la même étude. L'enquête de METR du 11 mai 2026 auprès de 349 collaborateurs techniques trouve une augmentation de valeur autodéclarée médiane de 1,4 à 2 fois, et ajoute elle-même qu'il y a des raisons de douter de cet ordre de grandeur.

La courbe de tendance mérite la même réserve. L'accélération vers un temps de doublement de 88,6 jours est mesurée sur une année. La série révisée contient quatorze modèles là où la précédente en comptait trente-trois. Les intervalles de confiance se sont resserrés mais restent larges: la borne supérieure pour le meilleur modèle se situe toujours au-delà du double de l'estimation ponctuelle. METR écrit lui-même que le nombre de tâches que la génération la plus

récente ne maîtrise pas est devenu petit, et que l’étalon menace donc de saturer par le haut. Une extrapolation sur une telle base est une hypothèse, pas une prévision.

Les chiffres d’observation avancés comme alternative mesurent eux aussi autre chose que ce qu’ils semblent mesurer. Qu’environ 4% des commits sur GitHub seraient écrits par un agent est un chiffre d’adoption. Il ne dit rien de la qualité, du volume de reprise ou du gain de temps net. Quelques participants à l’essai de METR ont remarqué qu’avec des agents ils abordaient d’autres tâches que sans, et que la qualité du résultat final différait entre les deux bras. Si la tâche change, une comparaison de temps ne mesure plus rien.

Pour finir, une tension que le secteur n’a pas encore réglée. Les règles à haut risque ont été reportées parce que les normes manquaient. À partir de décembre 2027, ces mêmes normes sont la mesure, avec une présomption de conformité comme effet juridique. Elles sont maintenant achevées sous une procédure d’exception qui saute le vote formel, et au sein du JTC 21 des objections ont été formulées ouvertement contre cela, avec une demande de révision de la mesure. Une norme qui naît par une voie raccourcie et qui décide ensuite si un produit peut légalement être sur le marché est une invitation à plaider. Qui bâtit en 2027 sa conformité sur ces normes bâtit sur un fondement dont la discussion est encore en cours.

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

À propos de Venture Campus

Venture Campus est un bureau de conseil belge en financement de l’innovation. Il prépare les demandes, les rédige et les introduit avec le client auprès des instruments flamands, fédéraux et européens.

Les critères d’évaluation sont neutres sur le plan sectoriel. Tous les secteurs, produits et services sont éligibles. Le client apporte la technologie, le domaine et le plan. Venture Campus apporte le régime, ses critères et le regard de l’évaluateur, appliqué avant l’introduction.

Pour le travail décrit ici, construire de la connaissance et lever l’incertitude technique, l’instrument O&O du VLAIO est en général le canal indiqué: 25 à 45% des coûts acceptés pour les projets de développement, 25 à 60% pour les projets de recherche, avec un maximum de 3 millions d’euros. Le VLAIO finance la connaissance et le risque technique, pas des produits. Les pourcentages valent à la date sur la couverture.

Sur la plupart des missions vaut le no cure, no pay: vous payez quand l’aide est octroyée. Cela ne vaut pas pour la Strategische Transformatiesteun et ne décrit pas les trajets fiscaux. L’analyse préalable est gratuite, tout comme le subsidiescan de six questions. S’il n’y a pas de cas réaliste, le bureau le dit.

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

SOURCES

Règlement (UE) 2024/1689 (règlement sur l’IA), articles 40 et 50 · Conseil de l’Union européenne, communiqué sur l’approbation définitive de l’omnibus numérique sur l’IA, 29 juin 2026 · Conseil de l’UE et Parlement européen, accord provisoire sur l’omnibus numérique sur l’IA, 7 mai 2026; approbation du Parlement européen 16 juin 2026 · Commission européenne, proposition de règlement simplifiant la mise en oeuvre des règles harmonisées en matière d’IA (omnibus numérique sur l’IA), 19 novembre 2025 · CEN-CENELEC, Update on CEN and CENELEC’s Decision to Accelerate the Development of Standards for Artificial Intelligence, 23 octobre 2025 · CEN-CLC/JTC 21, demande de normalisation M/593 et amendement M/613; projet de norme prEN 18286, enquête publique à partir du 30 octobre 2025 · METR, Measuring AI Ability to Complete Long Tasks, 19 mars 2025 (Kwa et al., arXiv:2503.14499) · METR, Time Horizon 1.1, 29 janvier 2026 · METR, Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity, 10 juillet 2025 (Becker, Rush, Barnes et Rein, arXiv:2507.09089) · METR, We are Changing our Developer Productivity Experiment Design, 24 février 2026 · METR, Measuring the Self-Reported Impact of Early-2026 AI on Technical Worker Productivity, 11 mai 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., réannotation de MMLU; Aleithan et al., audit de SWE-bench · Eriksson et al., méta-revue sur le benchmarking de l’IA, Centre commun de recherche de la Commission européenne, 2025 · Cantero Gamito, From Consensus to Exceptionality: What the EU’s AI Standards Crisis Reveals About Delegated Technical Governance, 28 novembre 2025