

Version en ligne : aipm-standard.com/aipm/cas/un-echec-sur-trois · édition du 30/09/2026. Le statut de preuve peut évoluer : la page en ligne fait foi.

Cas approfondi

À quatre points de l'humain, et un échec sur trois

Le rapport annuel de Stanford mesure les agents sur six bancs d'essai et constate qu'ils approchent la performance humaine — quatre points d'écart sur la navigation web, six sur le poste de travail. La même page ajoute qu'ils « échouent environ une fois sur trois ». Les deux phrases sont vraies, et c'est tout le problème de l'autonomie.

Toute la banque de cas (aipm-standard.com/aipm/cas) · La règle de preuve (aipm-standard.com/aipm/methode-de-preuve)

Statut de preuve : Confirmé opérationnel. Ce statut n'est pas une opinion sur l'entreprise ou sur la technologie : il décrit ce que les sources consultées permettent d'affirmer, et rien de plus. La règle qui l'attribue est publiée sur la page Méthode de preuve (aipm-standard.com/aipm/methode-de-preuve). Il porte sur le constat central de la fiche ; les chiffres cités à l'intérieur conservent le statut de leur propre source, et ceux qui proviennent du seul éditeur restent « annoncés », y compris dans une fiche confirmée.

Le contexte

Qui, quand, quel système

L'AI Index de Stanford est publié chaque année par le Human-Centered AI Institute. Ce n'est ni un éditeur de modèles, ni un cabinet vendant du conseil en IA : c'est un travail académique qui agrège des mesures publiées et les présente avec leurs protocoles. Dans une littérature dominée par les communications de fournisseurs, c'est une des rares sources sans intérêt au résultat.

Le chapitre 2 de l'édition 2026, consacré à la performance technique, réunit les mesures d'agents sur six bancs d'essai. C'est la vue la plus large disponible publiquement sur ce que les agents savent faire.

La décision, et qui l'a prise

La Directive de fait

Aucune décision d'entreprise n'est en jeu ici : c'est une mesure. Mais elle porte sur une décision que des milliers d'organisations prennent en ce moment — confier ou non une tâche à un agent sans supervision — et elle en donne les deux faces à la même page.

La progression est réelle et rapide. GAIA passe de 20 % en janvier 2025 à 74,5 % en septembre 2025. WebArena passe d'environ 15 % en 2023 à 74,3 % début 2026. MLE-bench, sur des tâches d'ingénierie de l'apprentissage automatique, passe de 17 % en 2024 à 64,4 % en 2026. Cybench, sur des tâches de cybersécurité, passe de 15 % à 93 % sur la même période.

Ce que le système a réellement fait

Les écarts à l'humain sont désormais minces. Sur WebArena, 74,3 % contre 78,2 % pour la référence humaine : quatre points. Sur OSWorld, 66,3 % contre environ 72 % pour des étudiants en informatique : six points. Sur GAIA, l'écart reste de 17,5 points, la référence humaine étant à 92 %.

Et pourtant le rapport écrit que les agents « échouent environ une fois sur trois ». Ce n'est pas une contradiction : un écart de quatre points sur une moyenne cohabite parfaitement avec un taux d'échec d'un tiers, dès lors que l'humain de référence échoue lui aussi souvent. Ce que la proximité à l'humain ne dit pas, c'est le niveau absolu — et le niveau absolu est celui d'un collaborateur qui rate une tâche sur trois.

L'état de la preuve

Ce que les sources établissent, et ce qu'elles n'établissent pas

Ce que la source établit

- Les six séries de mesures sont agrégées par une source académique sans intérêt commercial au résultat
- Les protocoles sont nommés et les tailles d'échantillon publiées : OSWorld teste 369 tâches en environnement contrôlé
- WebArena vérifie l'état final du site plutôt que la trace des actions — c'est une vérification par les effets, pas par le récit
- Le rapport publie ses propres limites : les problèmes de style concours sont plus structurés que le travail ouvert de la plupart des métiers

Ce que la source n'établit pas

- Ce que ces taux deviennent sur des tâches d'entreprise réelles, non structurées
- Comment se répartissent les échecs : une tâche difficile ratée systématiquement, ou un aléa sur toutes
- Le coût d'un échec — un banc d'essai compte les réussites, il ne pèse pas les conséquences
- Si la référence humaine est comparable : des étudiants en informatique ne sont pas les professionnels dont on parle de remplacer le travail

Sources consultées : Stanford HAI — AI Index 2026, chapitre 2, Performance technique

Ce que ce cas dit du référentiel

Ce cas donne aux **quatre seuils** d'AIPM leur assise chiffrée. Un système qui réussit deux fois sur trois n'est pas un système à superviser par exception : c'est un système à superviser par défaut. La question n'est pas « peut-on lui faire confiance » mais « combien coûte le tiers qui rate », et cette question se pose tâche par tâche, pas modèle par modèle.

Il éclaire aussi la **maturité à deux axes**. Sur l'axe technique, la réponse est publique et bonne : les agents progressent vite. Sur l'axe organisationnel, rien de ce qui précède ne dit si *votre* organisation sait absorber une erreur sur trois — détecter, corriger, reprendre. C'est le seul axe sur lequel un banc d'essai ne vous renseignera jamais.

Enfin, la vérification par l'état final plutôt que par la trace, telle que WebArena la pratique, est exactement la **preuve exécutable** du Standard. C'est une convergence utile : la meilleure pratique d'évaluation académique et la règle de preuve d'AIPM disent la même chose.

La consigne

Ce qu'un lecteur fait différemment après avoir lu cette fiche

Raisonnez en tâches, jamais en modèles. « Cet agent est bon » n'a pas de sens ; « cet agent réussit huit fois sur dix sur ce type de dossier, mesuré sur vingt cas » en a un. C'est la seule phrase qui permet de fixer un seuil.

Chiffrez le coût d'un échec avant de chiffrer le gain d'une réussite. Une tâche où l'erreur se détecte et se reprend en deux minutes supporte un tiers d'échecs. Une tâche où l'erreur part chez un client n'en supporte aucun. Le même taux, deux décisions opposées.

Mesurez votre reprise, pas seulement votre réussite. Le nombre utile n'est pas le taux d'échec de l'agent : c'est le temps qu'il vous faut pour repérer un échec et le rattraper. Une organisation qui rattrape en dix minutes peut déléguer beaucoup plus qu'une organisation qui découvre l'erreur au trimestre.

Ce qui invaliderait cette lecture. Une édition suivante du même rapport montrant les taux d'échec en chute nette sur des tâches non structurées, ou une évaluation portant sur du travail d'entreprise réel plutôt que sur des bancs d'essai. La consigne de raisonner par tâche resterait ; le chiffre d'un échec sur trois, lui, serait périmé.

Cette fiche suit le gabarit des cas approfondis d'AIPM : huit questions, toujours les mêmes. Voir les autres fiches (aipm-standard.com/aipm/cas-approfondis).

© 2026 DYONYSOS — Tous droits réservés · Bexo® est une marque déposée de DYONYSOS · aipm-standard.com. Reproduction interdite sans autorisation, hors courtes citations avec mention de la source (aipm-standard.com).