SPOTLAB / ANALYSE P04

Benchmarks LLM 2026 : pourquoi le premier du classement n'est pas toujours le meilleur choix pour l'entreprise

Une revue critique des benchmarks et une méthode de décision pour les entreprises

Analyse / P04

Novembre 2026

Scores datés des 27–28 octobre 2026 · sources consultées au 2 novembre 2026

ANALYSE P04

Sommaire

01	Trois enseignements à retenir	4
02	Ce que montrent réellement les chiffres d'octobre 2026	5
03	Pourquoi les principaux benchmarks ne répondent pas à la même question	7
04	La méthode Spotwork.AI : passer du classement à la décision	8
05	Exemple : sélectionner un modèle pour une étude stratégique	9
06	Ce que les décideurs doivent retenir	9
07	Analyse approfondie : mettre les benchmarks à l'épreuve	10
08	Construire un protocole professionnel reproductible	12
09	Exemple approfondi : rapport stratégique versus contrôle réglementaire	12
10	Méthode et références	13

FICHE DE L'ANALYSE

ÉDITEUR
Spotwork.AI — Spotlab, Research & Insights

RÉFÉRENCE
P04

TYPE
Analyse · revue comparative des évaluations publiées

COLLECTION
Modèles & architectures IA

SCORES
Datés des 27–28 octobre 2026

SOURCES CONSULTÉES
2 novembre 2026

PUBLICATIONS LIÉES
I0-01 · P05 · P06

> *Un benchmark compare des performances sur un protocole donné. Son résultat doit être remis en perspective avec les cas d'usage, les risques et les exigences propres à l'entreprise.*
>
> CE QU'IL FAUT RETENIR

INTRODUCTION

Un classement qui change en vingt-quatre heures

Les classements de modèles d'intelligence artificielle évoluent vite. Le 27 octobre 2026, Claude Opus 5.5 occupait la première place de l'Artificial Analysis Intelligence Index v5 avec un score de 58. Dès le 28 octobre, Gemini 4 Carbon le dépassait avec un score de 59, alors que son accès restait limité à certains utilisateurs [1][2].

27 OCTOBRE 2026 · LANCEMENT V5

58

Claude Opus 5.5 (max)
Premier de l'indice au lancement de la v5

28 OCTOBRE 2026 · ACTUALISATION

59

Gemini 4 Carbon (high)
Nouveau premier ; accès initialement limité à certains utilisateurs

Artificial Analysis Intelligence Index v5 : scores publiés à deux dates différentes. Il ne s'agit pas de tests Spotwork.AI.

Cette évolution n'invalide pas les benchmarks : elle rappelle ce qu'ils sont. Une mesure datée, appliquée à des tâches définies, qui aide à sélectionner des candidats — pas à décider seule quel modèle utiliser dans un processus professionnel.

Nous avons confronté les approches d'Artificial Analysis, de Stanford, d'Arena et de METR pour identifier ce que leurs résultats permettent réellement d'affirmer, puis construit une grille d'arbitrage directement utilisable par les entreprises. Cette publication est une revue de sources tierces ; Spotwork.AI n'a pas effectué de benchmark expérimental pour cette étude.

CONTRÔLE AU 2 NOVEMBRE 2026

Gemini 4 Carbon (high) reste en tête du classement Artificial Analysis avec un score de 59.

01 Trois enseignements à retenir

1

Un classement n'est fiable que dans son périmètre.

La suite Artificial Analysis v5 combine neuf évaluations, en anglais principalement, couvrant quatre catégories pondérées : travail de connaissance, code, sciences et capacités générales. Elle ne constitue pas un test universel de tous les documents professionnels, langues et contraintes métier [3].

2

La méthode peut changer le résultat.

La version v5 s'appuie à 60 % sur des jeux privés ou des réponses non divulguées. Elle remplace ou modifie plusieurs évaluations ; un écart entre v4 et v5 peut donc provenir de l'instrument de mesure plutôt que du modèle seul [1].

3

La performance utile se juge sur le travail accepté.

Le choix dépend de la qualité requise, des erreurs critiques, de la confidentialité, du coût des reprises et de la maîtrise des résultats. Un bon score général sert à présélectionner, mais une décision d'entreprise réclame des tests comparables sur ses propres cas.

Composition de l'Intelligence Index v5

Quatre familles pondérées à 25 % chacune [3]



Neuf évaluations au total · suite principalement textuelle et en anglais · 60 % de l'indice reposent sur des jeux privés ou des réponses non divulguées.

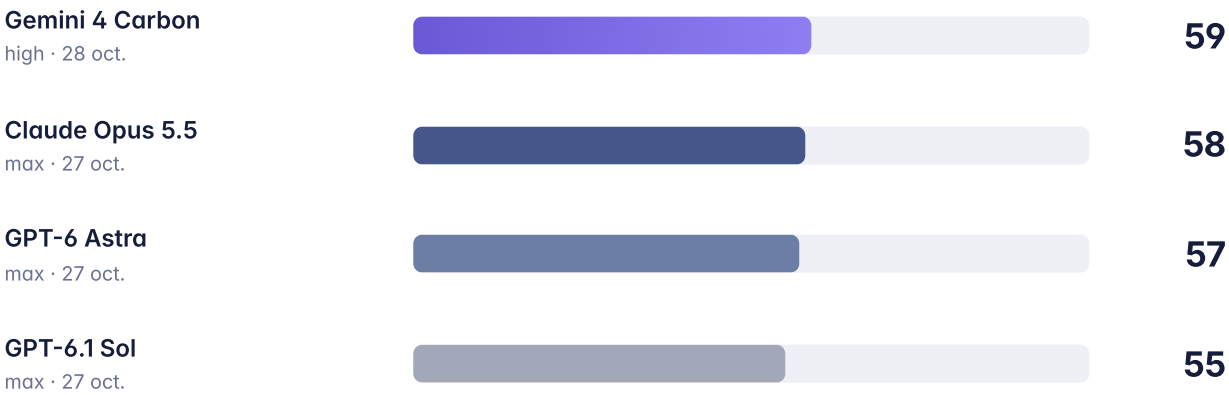
02 Ce que montrent réellement les chiffres d'octobre 2026

Modèle / configuration	Indice AA v5	Coût par tâche de l'indice	Lecture correcte
Gemini 4 Carbon — high	**59**	7,76 $*	Devient n° 1 le 28 octobre ; déploiement initial limité
Claude Opus 5.5 — max	**58**	10,04 $	Leader lors du lancement de v5 le 27 octobre
GPT-6 Astra — max	**57**	6,24 $	Score proche dans ce protocole précis
GPT-6.1 Sol — max	**55**	1,27 $	Coût par tâche inférieur pour une note agrégée proche

* Tarif standard ; Artificial Analysis signale une remise de lancement de 50 % sans date de fin annoncée [2].

Sources : communications Artificial Analysis des 27 et 28 octobre 2026 [1][2]. Scores et coûts fournis par l'éditeur de l'indice ; les dates et configurations sont essentielles. Le coût par tâche de l'indice n'est pas le prix d'une étude client et n'inclut pas nécessairement les coûts de revue métier.

Scores sur une échelle de 0 à 100 : des écarts de quelques points — INDICE AA V5



02 / CE QUE MONTRENT RÉELLEMENT LES CHIFFRES — SUITE

La disponibilité et la configuration comptent également

Le classement du 28 octobre illustre aussi une autre limite : une position élevée ne signifie pas qu'un modèle soit immédiatement accessible aux mêmes conditions pour tous les clients. Disponibilité, contrats, régions, performances en français et outils autorisés doivent être examinés séparément.

Artificial Analysis indique que Claude Opus 5.5 passe de 35 points en effort faible à 58 en effort maximal, avec un coût par tâche de l'indice passant de 0,98 à 10,04 dollars [1]. Ce n'est pas un argument pour choisir systématiquement la configuration la plus coûteuse : c'est la preuve qu'il faut comparer les paramètres autant que les noms des modèles.

EFFORT FAIBLE

35 points

0,98 $ par tâche de l'indice

EFFORT MAXIMAL

58 points

10,04 $ par tâche de l'indice

Claude Opus 5.5 · Artificial Analysis Intelligence Index v5, annonce du 27 octobre 2026. Résultat rapporté par l'éditeur de l'indice, pas un barème de prix universel.

> *Comparer « Claude Opus 5.5 » sans préciser faible ou maximal revient à comparer plusieurs configurations sous une même étiquette.*
>
> 07 / ANALYSE APPROFONDIE

03 Pourquoi les principaux benchmarks ne répondent pas à la même question

Évaluation	Ce qu'elle aide à apprécier	Ce qu'elle ne prouve pas à elle seule
Artificial Analysis Intelligence Index	Capacité agrégée sur neuf évaluations pondérées	Meilleur modèle pour un dossier client spécifique
Arena	Préférences humaines dans des comparaisons de réponses	Exactitude juridique, financière ou documentaire
SWE-bench et évaluations de code	Résolution de tâches logicielles sous un protocole donné	Qualité d'un rapport ou sécurité d'un workflow métier
METR Time Horizon	Probabilité de réussir des tâches en fonction de leur difficulté estimée en temps humain	Temps réel d'autonomie d'un agent, tous métiers confondus

Les classements de préférence d'Arena reposent sur des comparaisons par paires [5]. Les mesures de METR concernent principalement des tâches logicielles, de machine learning et de cybersécurité ; METR précise lui-même que son time horizon représente une difficulté estimée, pas une durée d'autonomie observée [6].

La qualité des jeux de tests est elle aussi un risque. Stanford documente des évaluations qui se saturent et des questions invalides [4]. OpenAI a publié des critiques détaillées de SWE-bench Verified et SWE-bench Pro [7][8] ; ces analyses d'un fournisseur de modèles sont utiles, sans devoir être prises comme une évaluation indépendante de toute la recherche.

STANFORD AI INDEX 2026

Des évaluations conçues pour durer des années saturent en quelques mois ; des questions invalides sont documentées.

OPENAI · FÉVRIER 2026

Critique de SWE-bench Verified : défauts de conception des tests et contamination.

OPENAI · JUILLET 2026

Critique de SWE-bench Pro : une part importante de tâches jugées défectueuses.

04 La méthode Spotwork.AI : passer du classement à la décision

Nous proposons un parcours d'arbitrage en quatre étapes, indépendant des fournisseurs, commun à notre guide de sélection « Comment choisir un modèle d'IA pour son entreprise ? Six critères et une méthode de décision ». Il s'agit d'un cadre de travail à éprouver, pas d'un score de fiabilité certifié.

Étape	Question	Preuve minimale
1. Cadrer	Quel travail, quel risque et quel résultat accepté ?	Cas métier, critères bloquants, référence de performance
2. Écarter (porte 1)	Parmi les candidats présélectionnés à l'aide des benchmarks, lesquels respectent les exigences bloquantes ?	Sources récentes, versions et variantes ; décision sécurité et confidentialité
3. Comparer (porte 2)	Quels candidats admissibles réussissent les mêmes cas dans notre contexte ?	Qualité, sources, stabilité, reprises et coût complet
4. Décider et réévaluer	Quelle configuration retenir et comment la faire évoluer ?	Décision tracée, seuils, supervision et non-régression

Les six critères de sélection définis dans « Comment choisir un modèle d'IA pour son entreprise ? Six critères et une méthode de décision » doivent rester visibles dans l'arbitrage : adéquation au cas d'usage, qualité vérifiable, sécurité, confidentialité et droits, coût complet, rapidité et stabilité, indépendance et évolutivité. Les exigences critiques — confidentialité, droit d'accès, erreurs bloquantes — s'appliquent avant tout calcul de score moyen.

1 · Adéquation | 2 · Qualité vérifiable | 3 · Sécurité, confidentialité, droits | 4 · Coût complet | 5 · Rapidité et stabilité | 6 · Indépendance

UNITÉ ÉCONOMIQUE

Pour comparer les options à résultat égal, une unité économique utile est le coût complet par livrable accepté : coût des modèles, des outils et des reprises, du contrôle humain et de l'exploitation imputable, rapporté aux livrables réellement approuvés. Cette convention requiert un périmètre comptable homogène ; elle ne se déduit pas d'un coût par tâche publié dans un benchmark (voir « Combien coûte réellement l'IA générative en entreprise ? Du token au livrable accepté »).

05 Exemple : sélectionner un modèle pour une étude stratégique

SCÉNARIO ILLUSTRATIF non résultat de tests Spotwork.AI

Une équipe doit produire une étude sectorielle couvrant dix acteurs, avec sources datées, tableaux comparatifs et recommandations.

Les benchmarks généraux permettent de retenir plusieurs candidats. Le test métier doit ensuite vérifier : la présence des dix acteurs, la fidélité des références, la cohérence des chiffres et des conclusions, les droits sur les données, les reprises et le délai jusqu'à approbation. Un document agréable à lire mais contenant une source inventée est rejeté, même si le modèle arrive premier au classement global.

Trois décisions deviennent possibles :

GO — DÉPLOYER

si la qualité et les contrôles sont tenus à coût acceptable

CORRECTION PUIS RETEST — ADAPTER

les instructions, les sources ou la revue si les erreurs sont corrigeables

NO-GO — ÉCARTER

le candidat si des défauts critiques persistent

06 Ce que les décideurs doivent retenir

La meilleure stratégie n'est pas de sélectionner durablement le « gagnant » d'un palmarès. Elle consiste à savoir identifier, tester et remplacer les capacités d'IA selon le travail à accomplir.

Concrètement, pour l'acheteur : présélectionner à partir des classements, décider sur un test métier, et prévoir dès le contrat la possibilité de changer de modèle sans perdre ses méthodes, ses contrôles et ses connaissances.

07 Analyse approfondie : mettre les benchmarks à l'épreuve

Le problème des scores composites

Une moyenne agrégée peut masquer des écarts importants entre catégories. La v5 d'Artificial Analysis affecte 25 % à chacune de ses quatre familles, mais leur composition interne varie : par exemple, la famille Knowledge Work couvre des livrables de fichiers et des tâches d'automatisation, tandis que le code inclut des modifications de dépôts et des opérations en terminal [3]. Le score final est donc le résultat d'une politique de pondération explicite ; il ne représente pas un marché universel du travail professionnel.

Pour une banque ou un cabinet de conseil, la pondération pertinente pourrait différer sensiblement : conformité des informations et sources autorisées peuvent constituer des contraintes éliminatoires plutôt que quelques points supplémentaires sur une moyenne. Le diagnostic commence donc par une cartographie des tâches, et non par un top dix de modèles.

La disponibilité et la configuration comme variables de décision

À un instant donné, un modèle peut ne pas être accessible via l'API recherchée, dans une région ou sous des conditions contractuelles acceptables. Le cas de Gemini 4 Carbon, leader de l'indice au 28 octobre mais initialement limité à certains utilisateurs, montre qu'un classement de capacités ne remplace pas une analyse d'éligibilité au déploiement.

Le réglage de l'effort de raisonnement crée une seconde variation : comparer « Claude Opus 5.5 » sans préciser faible ou maximal revient à comparer plusieurs configurations sous une même étiquette. Dans l'annonce v5, son score et son coût par tâche évoluent de 35 / 0,98 $ à 58 / 10,04 $ lorsque l'effort passe de faible à maximal [1]. Il s'agit d'un résultat rapporté par Artificial Analysis sur son indice, pas d'un barème de prix universel.

> *Le diagnostic commence par une cartographie des tâches, et non par un top dix de modèles.*
>
> 07 / ANALYSE APPROFONDIE

07 / ANALYSE APPROFONDIE — SUITE

La robustesse des méthodes d'évaluation

Quatre risques doivent être documentés lorsqu'un benchmark est utilisé comme preuve.

RISQUE 1

Contamination et préparation

Un jeu de tests public peut devenir connu, directement ou indirectement, dans les corpus d'entraînement ou de réglage. Des données privées peuvent réduire ce risque, sans garantir que le contenu évalué soit représentatif du travail métier.

RISQUE 2

Validité des cas

Un test peut exclure une bonne réponse, contenir une question ambiguë ou une contrainte contradictoire. Les travaux critiques de Stanford et les publications d'OpenAI sur SWE-bench montrent pourquoi le protocole et le taux de cas contestables doivent être étudiés [4][7][8].

RISQUE 3

Évaluateur

Certaines tâches reposent sur un test déterministe, d'autres sur une préférence humaine ou un LLM évaluateur. La robustesse de la notation ne s'infère pas du nombre de questions : elle nécessite une validation indépendante et des cas de désaccord.

RISQUE 4

Conditions d'exécution

Le résultat dépend des outils, prompts, délais, variantes, limites de contexte, nombre d'essais et règles d'arrêt. Un changement majeur sur l'une de ces dimensions doit figurer dans le rapport de comparaison.

Trois niveaux de preuve qui ne doivent pas être fusionnés

A

Preuve externe

Un éditeur de benchmark publie un résultat, avec sa méthode, ses dates et son périmètre. Ce résultat peut justifier la présélection, mais ne démontre pas le fonctionnement de notre processus.

B

Preuve métier

Une équipe rejoue des cas représentatifs sous contrôle, documente les erreurs et calcule le coût total jusqu'à validation. Les résultats sont pertinents pour ce contexte et ne se généralisent pas automatiquement à toutes les entreprises.

C

Preuve de gouvernance

La direction démontre qu'elle peut autoriser, contrôler, actualiser et remplacer la configuration retenue, avec un responsable humain identifié. La bonne qualité d'une réponse ponctuelle ne remplace pas ces contrôles.

08 Construire un protocole professionnel reproductible

Pour chaque famille de livrable, il est utile de formaliser :

Élément du protocole	Règle concrète
Population des cas	Cas réels anonymisés ou synthétiques documentés ; représentativité explicite
Périmètre	Tâche, langue, données accessibles, version de référence, risques
Candidats	Modèles exacts, paramètres, fournisseurs d'inférence, date de test
Critères	Qualité, couverture, exactitude des sources, contraintes, latence, coût
Évaluations	Tests déterministes lorsque possible, revues humaines calibrées et évaluateur IA audité
Reprises	Nombre borné de répétitions selon risque, avec conservation des échecs
Décisions	Accepté, à corriger ou rejeté selon critères connus avant les essais
Audit	Traçabilité des éléments, des modifications et des approbations

EXEMPLE DE RÈGLE DE DÉCISION (ILLUSTRATIVE)

Un modèle qui échoue à une exigence de confidentialité est exclu, quel que soit son score moyen. Un modèle admissible peut ensuite être choisi selon son taux de livrables acceptés, son coût complet, sa latence et la stabilité observée sur plusieurs cas. Les seuils et pondérations sont décidés par l'organisation et doivent être justifiés.

09 Exemple approfondi : rapport stratégique versus contrôle réglementaire

RAPPORT STRATÉGIQUE

Les enjeux prioritaires peuvent être : couverture du périmètre, pluralité des sources, qualité de l'argumentation, cohérence des chiffres, qualité de la présentation et délai jusqu'à approbation.

CONTRÔLE RÉGLEMENTAIRE

La règle peut être différente : une source manquante ou non autorisée, une absence de traçabilité ou une décision automatique sans habilitation peut entraîner le rejet immédiat, même si le texte est convaincant. Les benchmarks généraux restent informatifs, mais ne définissent pas ces seuils métier.

Cela montre pourquoi la méthode d'arbitrage doit rester commune, tandis que les critères d'acceptation doivent être adaptés à la tâche et à son risque. Ce n'est pas une contradiction : c'est le principe même d'une évaluation professionnelle contextualisée.

10 Méthode et références

Nature de la publication : analyse documentaire comparative (Artificial Analysis, Stanford AI Index, Arena, METR, rapports critiques sur les benchmarks de code). Les chiffres sont attribués à leurs éditeurs ; aucune performance propre à Spotwork.AI n'a été mesurée. Les résultats et disponibilités étant très évolutifs, la photographie des scores est datée des 27–28 octobre 2026 ; le classement a été contrôlé le 2 novembre 2026. La méthodologie de l'article ne constitue pas une certification de modèles.

[1] Artificial Analysis, Intelligence Index v5 (27 octobre 2026).
artificialanalysis.ai/articles/artificial-analysis-intelligence-index-v5

[2] Artificial Analysis, Gemini 4 Carbon, new #1 (28 octobre 2026).
artificialanalysis.ai/articles/gemini-4-carbon-new-number-1

[3] Artificial Analysis, méthodologie de l'Intelligence Index.
artificialanalysis.ai/methodology/intelligence-benchmarking

[4] Stanford HAI, AI Index 2026 — Technical Performance.
hai.stanford.edu/ai-index/2026-ai-index-report/technical-performance

[5] Chiang et al., Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference (2024).
arxiv.org/abs/2403.04132

[6] METR, Time horizons.
metr.org/time-horizons/

[7] OpenAI, Why we no longer evaluate SWE-bench Verified (février 2026).
openai.com/index/why-we-no-longer-evaluate-swe-bench-verified/

[8] OpenAI, Separating signal from noise in coding evaluations — SWE-bench Pro (juillet 2026).
openai.com/index/separating-signal-from-noise-coding-evaluations/

NOTE

Cette sélection de liens officiels ou institutionnels ne constitue pas une validation exhaustive de chaque assertion. Les données de benchmark sont datées.

Spotwork.AI

L'IA professionnelle de confiance

Mieux travailler aujourd'hui,
construire votre intelligence de demain.

SPOTLAB · RESEARCH & INSIGHTS · ANALYSE P04

Benchmarks LLM 2026 : pourquoi le premier du classement n'est pas toujours le meilleur choix pour l'entreprise

Une revue critique des benchmarks et une méthode de décision pour les entreprises

© 2026 SPOTWORK.AI