N°7 IA NEWSLETTER

Septembre 2026

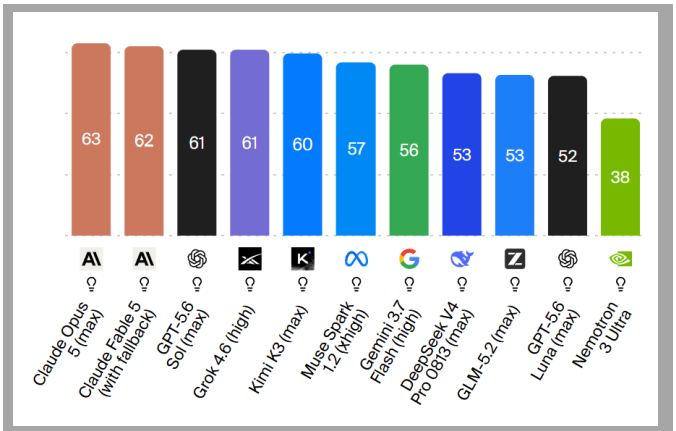
Top News

BENCHMARKER L'IA : COMMENT SUIVRE LES EVOLUTIONS ?

L'IA ne cesse d'évoluer : les modèles gagnent en puissance, se succèdent à un rythme effréné et se livrent une compétition permanente. Dans ce contexte, il devient de plus en plus difficile de mesurer leur performance réelle et, surtout, de **déterminer quel modèle privilégier selon les usages**.

Pour tenter d'y voir plus clair, **les benchmarks** se sont imposés comme des outils de référence, en comparant et classant les modèles.

Mais que mesurent réellement ces benchmarks ? Peut-on encore se fier à leurs résultats pour départager les modèles ? Et alors que les performances annoncées semblent progresser à grande vitesse, quelle est la part de réalité derrière ces scores ?



BENCHMARK INTELLIGENCE

Artificial Analysis – Septembre 2026

BENCHMARKER L'IA

Pourquoi Benchmarker l'IA ?

Face à la multiplication des modèles d'intelligence artificielle et leurs avancées toujours plus rapide, il est devenu essentiel de disposer de méthodes objectives pour **comparer leurs performances,** mais aussi pour d'identifier les domaines d'excellence des différents modèles.

En quoi ça consiste ?

Le benchmarking consiste à évaluer et comparer différents systèmes d'IA sur une série de tests standardisés afin de mesurer leurs capacités, leurs forces et leurs limites.

Les benchmarks s'appuient sur des jeux de données, des scénarios ou des questions de référence. Les modèles sont soumis aux mêmes épreuves, puis leurs résultats sont **mesurés selon des critères prédéfinis tels que l'exactitude, la pertinence, la rapidité, le coût ou encore la sécurité**.

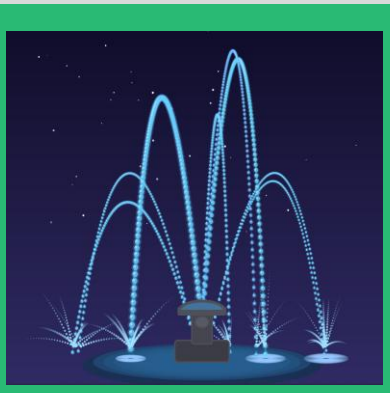
Qui réalise les benchmarks?

De nombreux acteurs réalisent ces évaluations : les laboratoires de recherche (OpenAI, Anthropic, Google DeepMind, Mistral AI, etc.), les organismes indépendants de recherche, les institutions publiques comme le **NIST** (*National Institute of Standards and Technology*), mais aussi les entreprises utilisatrices souhaitant sélectionner une solution.

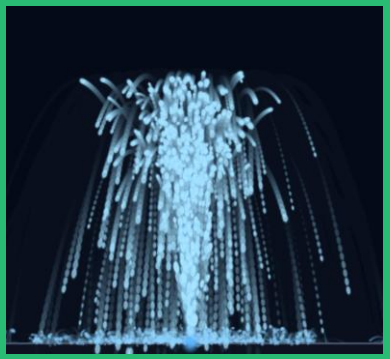
Artificial Analysis

Artificial Analysis est une plateforme de benchmark IA qui permet de comparer, en temps réel, les réponses générées par différents modèles d'intelligence artificielle à partir d'un même prompt.

Exemple d'instructions à suivre :
« Créez une animation p5.js représentant une fontaine à eau basée sur la physique des particules. Faites jaillir des gouttelettes d'eau depuis le bas et le centre du canevas ; elles doivent décrire une trajectoire en arc vers le haut avant de retomber, soumises à une gravité réaliste et à de légères variations de vitesse aléatoires. Les particules doivent éclabousser ou disparaître lorsqu'elles atteignent le sol ! »



Résultat par Claude Opus 4.5



Résultat par GPT-5.2

COMMENT FONCTIONNENT LES BENCHMARK ?

Constitution d'un jeu de tâches

- Questions de raisonnement
- Problèmes de code
- Suivis d'instructions
- Questions/réponses

Les sources de données sont très récentes pour éviter d'avoir un modèle déjà entraîné sur ces questions

Les différents modèles exécutent le même prompt avec les mêmes outils disponibles, avec la même sortie attendue

Notation et comparaison des sorties

Il existe **trois méthodes de notation :**

- Les réponses **vérifiables** (résultat numérique, QCM, résultat binaire)
- Les réponses **jugées par un agent IA spécialisé** (LLM-as-judge)
- Les réponses **jugées par un évaluateur humain**

LES BENCHMARS IA

NODEL	OVERALL	REASONING	CODING	AGENTIC CODING	MATHENATICS	DATA ANALYSIS	LANGUAGE	INSTRUCTION FOLLOWING	COST PER SUCCESSFUL TASK
Claude Fable S Max Effort	83.0	89.7	86.0	62.2	96.0	80.5	90.7	75.8	$1.439
GPT-S.6 Sol Max Effort	81.0	91.7	83.9	56.2	96.2	79.8	87.7	71.8	$0.438
GPT-S.S Thinking xHigh Effort	80.2	89.7	82.1	54.0	95.9	81.6	87.4	70.7	$0.435
Claude S Opus Thinking Max Effort	80.1	91.2	81.4	65.2	95.7	74.6	88.7	63.8	$0.699
Kimi K3 open	79.2	90.7	81.4	62.2	84.4	78.7	85.5	71.4	$0.348
GPT-S.4 Thinking xHigh Effort	78.0	88.1	77.5	53.8	94.1	79.3	82.6	70.2	$0.387
GPT-S.6 Terra Max Effort	77.9	90.6	78.2	54.9	94.9	79.3	82.9	64.6	$0.352
Gemini 3.1 Pro Preview High	77.0	84.0	76.5	44.1	91.0	78.5	85.4	79.1	$0.286

https://livebench.ai/#/ - Septembre 2026

LES CARACTÉRISTIQUES ESSENTIELLES

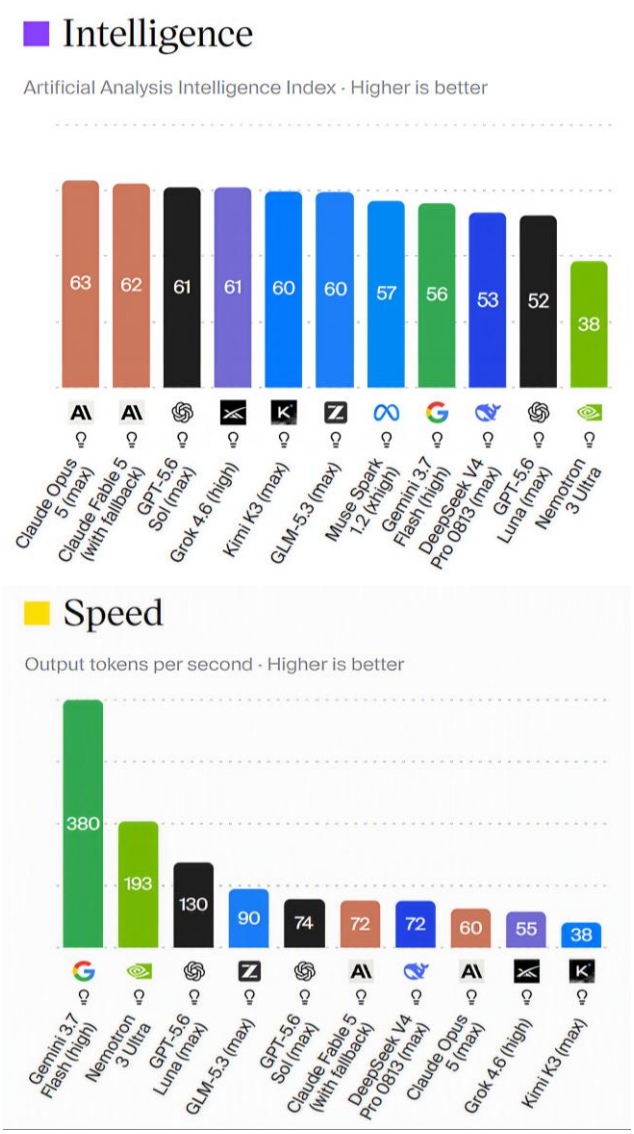
Une fois que l'ensemble des modèles a répondu à la totalité du jeu de questions, les résultats sont ventilés par type de problème puis synthétisés en une note globale.

Cette note agrège plusieurs évaluations distinctes (raisonnement scientifique, codage en conditions réelles, travail de bureau, etc…).
Trois catégories sont à distinguer pour comparer les modèles au-delà de la performance brute :

Intelligence : la note globale résume l'ensemble des tests (raisonnement, codage, mathématiques, etc.) en un seul indice. C'est le **critère le plus regardé, mais aussi le plus trompeur** si on l'utilise seul : un modèle **très intelligent peut rester inadapté à un usage si sa vitesse ou son coût ne suivent pas** (exemple : Fable 5 de Claude).

Rapidité : le **nombre de tokens générés par seconde**. Un critère indépendant de l'intelligence : deux modèles avec une note globale identique peuvent afficher des vitesses très différentes, ce qui devient déterminant pour tout usage nécessitant une réponse quasi instantanée (agent conversationnel, outil interactif).

Prix moyen selon la performance : un modèle moins cher au token mais qui échoue davantage peut revenir plus cher qu'un modèle plus fiable. Il faut donc **prendre en compte le prix moyen dépensé par tâches réussies.**



https://artificialanalysis.ai/models

LES LIMITES DES BENCHMARKS

CE QUE MESURENT LES MODÈLES & CE QU'ILS NE MESURENT PAS

Les benchmarks IA permettent de comparer les modèles sur des tâches et des critères standardisés, afin de mesurer leurs performances sur des capacités précises : raisonnement, connaissances, programmation ou compréhension du langage.

Mais ils **ne reflètent pas nécessairement leur capacité à réaliser efficacement des tâches complexes dans des situations réelles**. Un bon score ne garantit donc pas qu'une IA soit fiable, autonome ou performante dans un contexte professionnel.

SWE-bench

SWE-bench est un benchmark qui mesure la capacité des modèles IA à résoudre de vrais problèmes de programmation issus de projets GitHub. On soumet au modèle un code erroné. Il doit alors comprendre le problème, modifier le code et faire passer les tests.

SWE-bench Verified

La version Verified repose sur 500 problèmes vérifiés par des développeurs humains, offrant une mesure plus fiable des capacités de coding agentique.

SWE-bench lancer

SWE-Lancer évalue la capacité des IA à réaliser de vraies missions de développement informatique, issues de missions freelance sur Upwork.

En novembre 2025, une étude de l'Oxford Internet Institute portant sur 445 benchmarks majeurs a démontre que ces tests **manquent en réalité de rigueur scientifique** et ont **tendance à surestimer les performances des modèles.**

Sur SWE-bench Verified, les meilleurs modèles affichent environ 80 % de réussite sur des tâches de codage. Avec le modèle SWE-Lancer, qui évalue les mêmes modèles sur de vraies missions freelance, le taux de réussite tombe à 26,2 %

Face à ce constat de manque de rigueur scientifique sur de nombreux benchmarks, OpenAI **recommande désormais l'utilisation de SWE-bench Pro, une version renforcée du benchmark pour le code.** Les performances y chutent drastiquement, avec les meilleurs modèles qui plafonnent autour de 23 %, contre 80 % sur la version Verified. L'entreprise investit également dans GDPVal, un système d'évaluation où les tâches sont rédigées par des experts et notées par des évaluateurs humains. L'approche est plus coûteuse, mais garantit une meilleure fiabilité des résultats.

SOURCES :

- LiveBench, https://livebench.ai/#/
- Artificial Analysis, https://artificialanalysis.ai/models
- NIST, *Practices for Automated Benchmark Evaluations of Language Models* (2026)
- NIST, *Expanding the AI Evaluation Toolbox with Statistical Models* (2026)
- OECD, *On evaluating artificial intelligence systems: Competitions and benchmarks*
- NIST, *AI Technology Evaluation (AITE)*
- https://www.blogdumoderateur.com/benchmarks-ia-evaluations-modeles-crise-credibilite/