HUMANITY'S LAST EXAM · BENCHMARK IA

à un quiz sur les colibris.

48,4%

Ce que ce score ne mesure pas →

Le test

Humanity's Last Exam

2 500

questions

« À la frontière de la connaissance humaine »

positionnement officiel

Le benchmark de référence

pour évaluer si une IA peut faire de la science

Score record actuel : **Gemini 3 Deep Think · 48,4%**

Exemples

À quoi ressemblent les questions ?

> *« Combien de tendons appariés l'os sésamoïde d'un colibri soutient-il ? »*

> *« Combien de couleurs ont les allotropes du phosphore ? »*

« En quoi savoir ça aide qui que ce soit à découvrir quoi que ce soit ? »

— Chenru Duan, fondateur de Deep Principle

Petit rappel

La psychométrie sait ça depuis 1954.

Standards APA. Deux types de validité.

L'industrie de l'IA est en train de les redécouvrir.

01

Validité de contenu

La question :

« Le test couvre-t-il bien le domaine ? »

Verdict pour HLE :

Imbattable.

2 500 questions, des dizaines de disciplines, validation par des experts.

02

Validité prédictive

La question :

« Les scores prédisent-ils la performance réelle ? »

Verdict pour HLE :

On n'en sait rien.

Personne n'a montré qu'un modèle à 48% découvre plus qu'un modèle à 8%.

Un test peut être impeccable sur le premier critère et complètement vide sur le second.

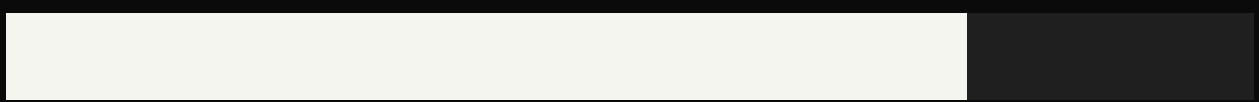
Benchmark sérieux n°1

FrontierScience

OpenAI · décembre 2025 · 700 questions

GPT-5.2

Questions type Olympiades



77%

Problèmes de recherche ouverts

25%

Le grand écart entre « je récite »

et « je cherche ».

Benchmark sérieux n°2

SDE

Scientific Discovery Evaluation · Deep Principle

1 125	8
tâches	projets de recherche réels

Les modèles top de tous les fournisseurs

OpenAI · Anthropic · xAI · DeepSeek

butent sur les mêmes questions difficiles.

Mêmes corpus.

Mêmes angles morts.

Benchmark sérieux n°3

LABBench2

FutureHouse · pipeline de recherche complet · ≈ 1 900 tâches

OK	Recherche bibliographique dans les brevets et essais cliniques
KO	Croisement de plusieurs bases de données
KO	Lecture et interprétation de figures denses dans des papiers

C'est exactement ce que fait

un chercheur la moitié de la journée.

La leçon

La métrique fabrique l'évidence.

> *« Quand une mesure devient un objectif, elle cesse d'être une bonne mesure. »*
>
> — Loi de Goodhart

Quand HLE devient le KPI de l'industrie,

on optimise pour HLE. Pas pour la science.

Tant qu'on mesurera

la science avec

des QCM de culture générale,

on aura des IA

championnes de QCM.

Pas des chercheurs.

Ce qu'on mesure, c'est ce qui s'améliore. Le reste meurt.

eudonia
accompagner pour réussir

On construit des outils IA

pour les pros.

Pas pour les remplacer.

Structurez l'analyse

de vos accompagnements

en quelques minutes.

eudonia.fr

eudonia
accompagner pour réussir