S3

SÉANCE 3 · LES CLASSEMENTS

Comparer des IA à l'aveugle

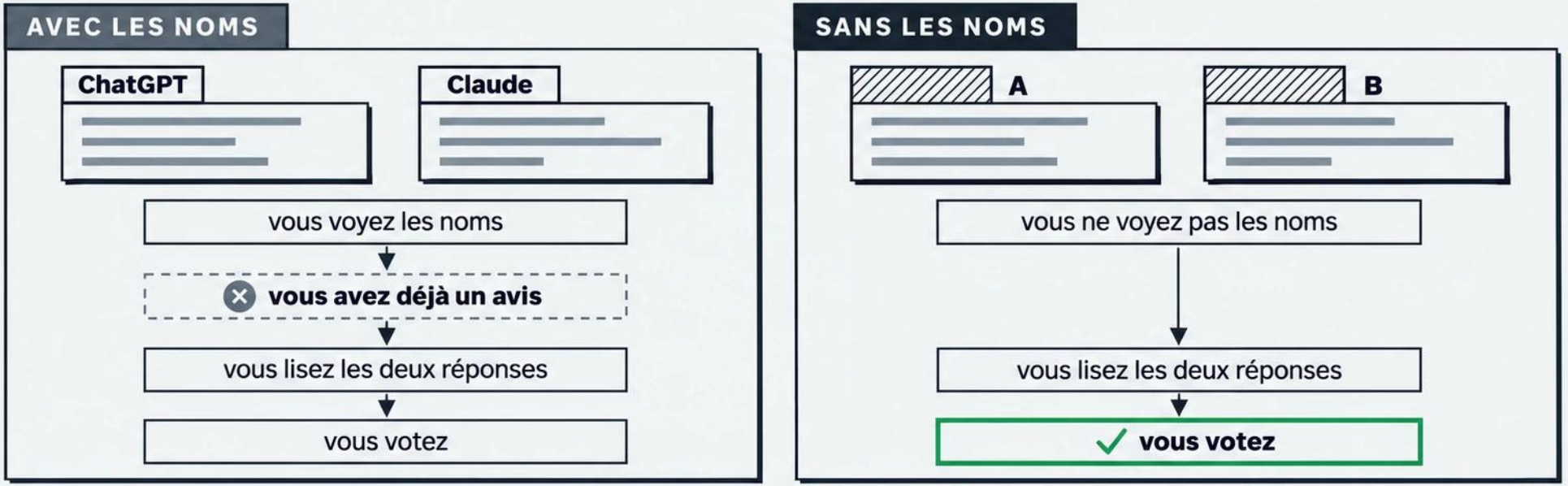
IFOSUP Wavre · 2026-2027

5XVTE · VEILLE TECHNOLOGIQUE · S3



? Si vous savez qu'une réponse vient de ChatGPT, la lisez-vous comme les autres ?

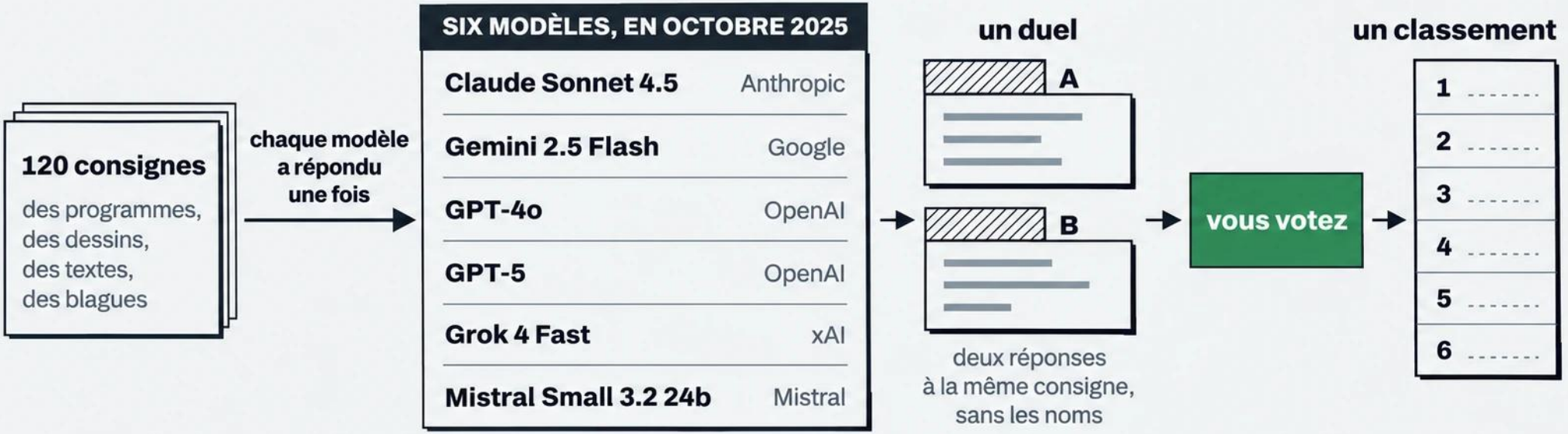
On cache les noms des IA, sinon vous avez déjà un avis avant de lire.



Arena, le site de classement du chapitre précédent, cache les noms et fait voter ses visiteurs. Vous allez faire la même chose avec la classe.

? LLM Face-Off est un site du professeur. Comment ce site classe-t-il des IA ?

Sur LLM Face-Off, des votes à l'aveugle classent six modèles sur les mêmes consignes.



Ces réponses ont un an, et les entreprises ont sorti d'autres modèles depuis. L'exercice ne cherche pas le meilleur modèle d'aujourd'hui. Il montre comment un classement se fabrique.

Vous allez voter dans vingt-cinq duels, sans savoir quelle IA a écrit quoi.

À VOUS : QUINZE MINUTES, CHACUN SUR SON ORDINATEUR

- ☐ Ouvrez `llm.choboai.com`
- ☐ Choisissez « 25 questions » et le groupe « 5XVTE ».
- ☐ Lisez la consigne, comparez les deux réponses, votez. `1` **A** `2` **B** `0` **égalité**
- ☐ Après chaque vote, notez sur une feuille un mot qui dit pourquoi.

« plus courte » **« le bouton marche »** **« plus drôle »** **« elle a refusé »**

Une consigne vous gêne ? Votez l'égalité et passez à la suivante.

Vous avez fini ? Faites une deuxième série de 25.

? Voici le classement du site le 23 novembre, avec 2 260 votes. Quel est le meilleur modèle ?

Au score Elo, GPT-5 est premier ; au taux de victoire, c'est Claude Sonnet 4.5.

`llm.choboai.com, le 23 novembre, tous les groupes : les trois premières lignes sur six`

RANG	MODÈLE	SCORE ELO	TAUX DE VICTOIRE	DUELS	V / D / E
1	GPT-5 LOW	1112	63.5%	748	407 / 205 / 136
2	Gemini 2.5 Flash HIGH	1056	57.2%	752	363 / 255 / 134
3	Claude Sonnet 4.5 LOW	1050	66.0%	759	447 / 204 / 108

1 **premier: GPT-5, 1112**
à chaque duel, le gagnant prend des points au perdant

2 **premier: Claude Sonnet 4.5, 66,0 %**
la part des duels gagnés, une égalité compte pour la moitié

V : victoires
D : défaites
E : égalités
LOW, HIGH : le réglage

Regardez la colonne « Score Elo », puis la colonne « Taux de victoire ». Elles sont calculées avec les mêmes votes, et Claude Sonnet 4.5 passe de la troisième place à la première.

? Les deux colonnes viennent des mêmes votes. Pourquoi n'ont-elles pas le même premier ?

Il y a deux premiers parce qu'il y a plusieurs façons de compter les mêmes votes.

2 260 votes
les mêmes pour les trois

LA FAÇON DE COMPTER	CONTRE QUI ON GAGNE	L'ORDRE DES DUELS	SON PREMIER
Le taux de victoire 447 victoires + 108 égalités ÷ 2, sur 759 duels	✗ ne compte pas	✗ ne compte pas	**Claude Sonnet 4.5** 66,0 %
Le score Elo le gagnant prend des points au perdant	✓ battre un mieux classé rapporte plus	✓ les derniers duels pèsent plus	**GPT-5** 1112
Bradley-Terry la formule d'Arena	✓ compte	✗ ne compte pas	**Claude Sonnet 4.5** 1104

ces nombres ne se comparent pas entre eux

Personne n'a triché : chaque calcul regarde autre chose dans les mêmes votes. Le score Elo vient du jeu d'échecs. Avec la formule d'Arena, GPT-5 est deuxième, à 1088.

? En plus de la façon de compter, qu'est-ce qui peut changer ce classement ?

Un classement dépend aussi des consignes, des réglages et de ceux qui votent.

LES QUESTIONS DU CHAPITRE PRÉCÉDENT	POUR LLM FACE-OFF
1 **Mesuré sur quoi ?**	120 consignes, une réponse par modèle
2 **Comparé à quoi ?**	six modèles, qui n'ont pas reçu le même réglage **effort bas, «low»** GPT-5, Claude Sonnet 4.5 — **effort élevé, «high»** Gemini 2.5 Flash, Grok 4 Fast — GPT-4o et Mistral : pas affiché
3 **Décidé par qui ?**	le professeur, pour les consignes et les réglages ; les votants, pour les gagnants groupe PROFS **premier: GPT-5** — groupe BAC3 INFO **premier: Claude Sonnet 4.5**

À vous, page Résultats : filtrez par votre groupe et comparez votre classement à celui de la classe

L'effort règle le temps que le modèle passe à réfléchir avant de répondre. Les groupes PROFS et **BAC3 INFO** n'ont pas le même premier, quelle que soit la façon de compter.

? Une entreprise annonce que son modèle est premier. Premier sur quel site, et à quelle date ?

Cinq sites donnent quatre premiers : ils ne mesurent ni la même chose, ni les mêmes modèles.

LE SITE	IL CLASSE D'APRÈS	SON PREMIER		MIS À JOUR
Arena	les votes à l'aveugle de ses visiteurs	**Gemini 4 Carbon** 1527		22 novembre
Artificial Analysis	neuf tests que le site fait passer	**Claude Fable 5.2** 65		pas de date
LiveBench	des travaux à rendre : présentations, code, applications	**Gemini 4 Neon** 87,4		11 novembre
Design Arena	des votes sur des sites, jeux, applis et diapositives	**Claude Opus 5.5** 1456	**le même premier**	23 novembre
compar:IA	des votes à l'aveugle, sur un site de l'État français	**Claude Opus 5.5** 1190		23 novembre

Ces résultats ont été relevés le 23 novembre, et le premier a peut-être changé.
Claude Fable 5.2, premier chez Artificial Analysis, est huitième sur Arena et absent de compar:IA.

? Sur Arena, le premier est-il beaucoup plus fort que le dixième ?

Le premier d'Arena gagne environ 54 duels sur 100 contre le dixième.

LE RANG le 22 novembre

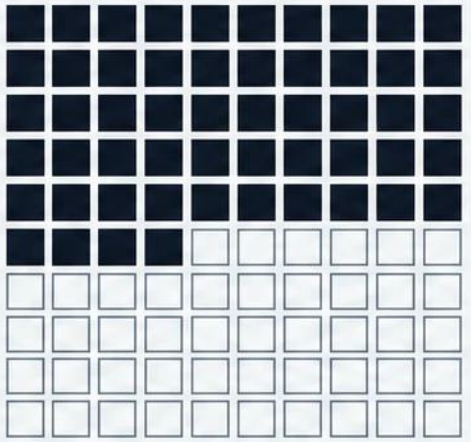
1	**Gemini 4 Carbon**	Google	1527
2	**Gemini 4 Argon**	Google	1520
3	**Gemini 4 Neon**	Google	1512
4	**Muse Wonder**	Meta	1512
5	**Claude Opus 5.5**	Anthropic	1507
...			
10	**Claude Opus 4.7**	Anthropic	1501

9,2 millions de votes, 427 modèles

26 points d'écart

se lit comme un score Elo

L'ÉCART



100 duels entre le premier et le dixième

■ gagnés par le premier : environ 54

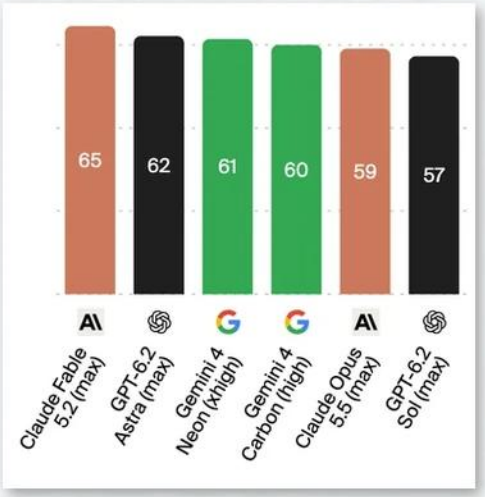
□ gagnés par le dixième : environ 46

Le rang impressionne plus que l'écart. Arena est une entreprise : elle vend des évaluations, entre autres à ceux qui font les modèles.

? Artificial Analysis donne une note à chaque modèle. Que vérifiez-vous avant de la citer ?

Avant de citer une note d'Artificial Analysis, regardez le réglage, la version et le prix.

artificialanalysis.ai, le 23 novembre : les six premiers



Intelligence Index
une seule note, tirée de neuf tests

1

1 Le réglage « max » est l'effort le plus élevé
votre IA gratuite peut répondre avec moins d'effort, ou être un autre modèle

2 La version du test
le modèle GLM-5.3 | 45 | version 5, le 27 octobre → | 32 | le même modèle, un autre test

3 Le prix moyen d'une tâche
Claude Fable 5.2, note 65 — 26,54 dollars
GPT-6.2 Sol, note 57 — 1,01 dollar
huit points de plus, 26 fois plus cher

Artificial Analysis est aussi une entreprise. Deux notes se comparent seulement dans la même version. Sur le graphique entier, dans le chapitre, cherchez l'IA que vous utilisez.

? Parmi vos mots, « le bouton marche » et « plus drôle » se discutent-ils de la même façon ?

Un critère se vérifie ou se juge, et on n'en discute pas de la même façon.

À vous : cinq minutes rangez vos mots dans ces deux colonnes :
ce sont vos critères, les choses que vous regardez dans une réponse

CE QUI SE VÉRIFIE

deux personnes qui regardent bien sont forcément d'accord

le programme marche
la réponse est juste
la consigne est suivie
elle fait moins de cent mots

vos mots .....
.....

CE QUI SE JUGE

deux personnes peuvent avoir raison toutes les deux

c'est beau
c'est clair
c'est drôle
c'est original

vos mots .....
.....

On ne discute pas de la même façon d'un bouton qui ne marche pas et d'une couleur qu'on n'aime pas.
Les deux sortes sont utiles : gardez cette feuille pour le chapitre suivant.