

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

Cas approfondi

Le même modèle, deux mesures, deux conclusions opposées

Sur le tableau de son éditeur, GPT-6 Astra écrase la génération précédente : 99,9 % contre 17,8 % sur un test de raisonnement — chiffre annoncé, non rejoué. Sur l'index d'un tiers indépendant, mesuré et publié, il est à égalité avec son concurrent direct. Ce qui est confirmé ici, c'est la divergence des deux récits, pas les scores de l'éditeur. C'est la leçon.

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

En septembre 2026, OpenAI publie l'annonce de GPT-6 Astra — page consultée le 10 — avec un tableau comparatif contre GPT-5.6 Sol, la génération précédente de la même maison. Les écarts affichés sont spectaculaires : ARC-AGI-3 à 99,9 % contre 17,8 %, ExploitBench à 100 % contre 78,5 %, FrontierMath Tier 4 à 97,6 % contre 83,0 %.

Le 9 septembre 2026, Artificial Analysis — un évaluateur indépendant qui rejoue ses propres tests — publie ses mesures. Son Intelligence Index place Astra à 53. Son Coding Agent Index le place à 62. Dans les deux cas, à égalité avec Claude Fable 5.1.

La décision, et qui l'a prise

La Directive de fait

Personne n'a menti et rien n'est contradictoire. L'éditeur a choisi ses tests et son point de comparaison : lui-même, six mois plus tôt. Le tiers a choisi les siens et son point de comparaison : le marché, aujourd'hui. Les deux choix sont défendables, et ils produisent deux récits inconciliables du même objet.

La décision qui compte est celle du lecteur : sur quel chiffre engage-t-il un budget ? C'est très exactement la décision qu'AIPM cherche à outiller.

Ce que le système a réellement fait

Une convergence mérite d'être relevée, parce qu'elle est rare et qu'elle vaut davantage que les écarts. Sur Terminal-Bench 4.0, l'éditeur annonce 57,9 % ; le tiers mesure 59 %. Deux protocoles, deux équipes, un écart d'un point : ce chiffre-là est solide, et il est le seul du lot à pouvoir être qualifié de confirmé.

Sur le reste, les périmètres ne se recouvrent pas : les tests du tableau de l'éditeur ne sont pas ceux de l'index du tiers. Il n'y a donc ni confirmation, ni contradiction — il y a deux ensembles disjoints présentés au même public dans la même semaine.

L'état de la preuve

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

Ce que la source établit

- Terminal-Bench 4.0 est mesuré deux fois, indépendamment : 57,9 % annoncé, 59 % constaté
- Les indices du tiers placent Astra à égalité avec son concurrent direct, pas devant
- Les tests du tableau de l'éditeur ne sont pas repris par le tiers, et réciproquement
- Le tiers publie son protocole et ses dates ; l'annonce publie ses scores

Ce que la source n'établit pas

- Si les écarts annoncés sur ARC-AGI-3, ExploitBench et FrontierMath sont reproductibles : aucun tiers ne les a rejoués à cette date
- Ce que ces tests prédisent du travail réellement payé en entreprise
- Si l'égalité mesurée par le tiers tiendra à la prochaine mise à jour de son index

Sources consultées : OpenAI — annonce de GPT-6 Astra · Artificial Analysis — mesures indépendantes, 9 septembre 2026

Ce que ce cas dit du référentiel

C'est le **principe 7** — l'interdiction de l'autorité empruntée — appliqué à un chiffre plutôt qu'à une institution. Un score n'emprunte pas seulement l'autorité de celui qui le publie : il emprunte celle du test, et celle du point de comparaison choisi. Trois emprunts, trois endroits où l'affirmation peut céder.

Le cas fournit aussi la règle la plus simple de toute la méthode de preuve d'AIPM : **un chiffre mesuré deux fois par deux parties qui ne partagent pas d'intérêt vaut mieux que dix chiffres mesurés une fois**. Terminal-Bench à 57,9 et 59 est plus utile que ARC-AGI-3 à 99,9.

La consigne

La consigne

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

Avant d'inscrire un chiffre de performance dans une note interne, posez trois questions dans cet ordre.

Qui a mesuré ? Si c'est le vendeur, le chiffre entre au statut « annoncé » et n'en sort que si quelqu'un d'autre le retrouve.

Contre quoi ? Un progrès contre soi-même n'est pas un avantage sur le marché. Exigez un point de comparaison externe, ou traitez le chiffre comme une communication.

Sur quoi ? Le test le plus proche du travail que vous payez l'emporte sur le test le plus impressionnant. Un score de raisonnement à 99,9 % ne dit rien de la capacité à finir un dossier.

Ce qui invaliderait cette lecture. Un tiers qui rejouerait ARC-AGI-3 et ExploitBench et retrouverait les scores annoncés. L'écart cesserait alors d'être un problème de méthode pour devenir une simple différence de couverture, et cette fiche perdrait l'essentiel de sa portée.

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).