

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

Cas approfondi

Le modèle plus rapide qui régresse là où le travail se paie

Sur le test le plus proche du travail professionnel réel, GPT-6 Astra perd environ 45 points Elo par rapport à la génération précédente. La cause mesurée : il utilise 24 tours par tâche à effort maximal, contre 45 pour la génération précédente et 60 pour Claude Fable 5.1 et Claude Opus 5. La vitesse a un prix, et il est chiffré.

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

La famille GDPval a été construite pour évaluer les modèles sur des tâches correspondant à du travail effectivement rémunéré, plutôt que sur des problèmes de concours. C'est, parmi les tests publics, celui dont le résultat se traduit le plus directement en décision d'entreprise.

Le 9 septembre 2026, Artificial Analysis publie ses mesures sur GPT-6 Astra. Sur GDPval-AA v2, l'évaluateur constate une baisse d'environ 45 points Elo par rapport à GPT-5.6 Sol — et l'attribue explicitement au fait que le modèle « utilise significativement moins de tours que d'autres modèles » : 24 par tâche à effort maximal, contre 45 pour GPT-5.6 Sol et 60 pour Claude Fable 5.1 et Claude Opus 5.

La décision, et qui l'a prise

La Directive de fait

Le compromis a manifestement été assumé côté conception. L'annonce de l'éditeur — chiffres annoncés, non rejoués par un tiers — met en avant la rapidité comme argument principal : environ 47 % de temps en moins par tâche sur OSWorld 2.0, 40 minutes en moyenne contre 75 pour la génération précédente, et un score supérieur — 72,6 % contre

65,7 %.

Sur OSWorld, faire moins de tours et aller plus vite améliore le résultat. Sur GDPval, la même économie le dégrade. Ce n'est pas une incohérence : c'est la signature d'un arbitrage qui gagne sur un type de tâche et perd sur un autre. L'organisation qui déploie hérite de l'arbitrage sans l'avoir choisi.

Ce que le système a réellement fait

Deux autres mesures du même relevé méritent d'être mises à côté, parce qu'elles vont dans le sens inverse et interdisent une lecture simpliste. Sur AA-Omniscience, le taux d'hallucination passe de 92 % à 51 %. Sur AA-Briefcase, le modèle gagne environ 90 points Elo.

L'évaluateur souligne que cette baisse ne se paie pas en exactitude, comme c'est souvent le cas : l'exactitude progresse de 4 points dans le même temps. Le même modèle, mesuré par la même équipe le même jour, progresse fortement sur deux axes et régresse sur celui qui ressemble le plus à du travail facturable.

L'état de la preuve

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

Ce que la source établit

- La baisse d'environ 45 points Elo sur GDPval-AA v2 est mesurée par un tiers indépendant
- La cause avancée est mesurée elle aussi : 24 tours à effort maximal, contre 45 et 60
- Le même relevé constate des progrès nets ailleurs : hallucinations de 92 % à 51 %, environ +90 Elo sur AA-Briefcase
- L'évaluateur précise que la baisse d'hallucination s'accompagne d'un gain d'exactitude de 4 points, et non d'une perte

Ce que la source n'établit pas

- Si la régression tient au nombre de tours ou à autre chose : la corrélation est constatée, le mécanisme n'est pas démontré
- Si un paramétrage utilisateur — autoriser davantage de tours — la ferait disparaître
- Ce que 45 points Elo représentent en résultat concret sur un dossier réel
- La position d'OpenAI sur ce point : l'annonce ne mentionne pas GDPval

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

Ce que ce cas dit du référentiel

Ce cas remplit une case que la banque de cas d'AIPM avait presque vide : **une régression mesurée**. Les cas disponibles décrivent presque toujours des gains, parce que ce sont les gains qu'on publie. Un référentiel qui n'aurait que des cas de progrès enseignerait à ne chercher que des progrès.

Il alimente directement les **quatre seuils**. Un seuil d'autonomie ne peut pas être fixé sur la performance moyenne d'un modèle : il doit l'être sur la performance de la classe de tâches concernée. Le même agent mérite un seuil différent selon qu'il pilote un poste de travail ou qu'il produit un livrable professionnel.

Il touche enfin la **maturité à deux axes** : une organisation peut être mature sur l'outillage et immature sur le choix des tâches confiées. C'est exactement le piège ici.

La consigne

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

Deux gestes.

Choisissez votre test avant de choisir votre modèle. Écrivez, en une phrase, la tâche que vous payez aujourd'hui à un humain et que vous envisagez de confier. Cherchez ensuite le test public qui lui ressemble le plus. C'est celui-là qui décide, pas l'index global.

Mesurez les tours, pas seulement le temps. Un agent qui finit en deux fois moins de temps parce qu'il fait deux fois moins d'essais n'est pas plus rapide : il abandonne plus tôt. Sur les tâches longues, le nombre d'itérations avant rendu est un indicateur de qualité, et il est mesurable chez vous, sans attendre un benchmark.

Ce qui invaliderait cette lecture. Une seconde mesure indépendante de GDPval qui ne retrouverait pas la régression, ou une démonstration que le nombre de tours est un paramètre réglable par l'utilisateur — auquel cas il ne s'agirait plus d'une limite du modèle mais d'un défaut de configuration, et la consigne changerait entièrement.

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).