Pourquoi Nous Avons Besoin de Benchmarks de Gouvernance Maintenant : Présentation de LNDT3.1 | Aeterna Claritas

Depuis plus d'une décennie, les benchmarks d'IA ont été notre boussole. Ils nous ont dit quand les machines pouvaient reconnaître les images mieux que les humains, quand elles pouvaient lire et répondre aux questions plus rapidement, quand elles pouvaient réussir des examens standardisés.

Les benchmarks comme ImageNet, GLUE, SuperGLUE, MMLU, et AGIEval sont devenus des jalons — cités dans les articles de recherche, mis en avant dans les communiqués de presse, célébrés dans les gros titres. Ils ont alimenté l'essor de l'apprentissage profond et des grands modèles de langage.

Mais voici le paradoxe : les benchmarks sont saturés plus rapidement que jamais. GPT-4 et DeepSeek surpassent les humains sur beaucoup d'entre eux. De nouveaux classements apparaissent chaque mois, mais chacun est rapidement "résolu". Les courbes de performance continuent de grimper, tandis que les évaluateurs humains sont laissés pour compte.

Bientôt, aucun jury humain ne sera capable de vérifier manuellement l'étendue de ce que produisent les modèles de pointe. Et cela signifie : nous devons repenser à quoi servent les benchmarks.

La Dimension Manquante : La Gouvernance

Les benchmarks traditionnels mesurent ce que l'IA peut faire.

Mais ils ne disent rien sur ce que l'IA devrait être autorisée à revendiquer.

C'est là que **LNDT3.1 | Aeterna Claritas** entre en scène.

La Clause Claritas

score_interne = le diagnostic interne, produit par l'IA ou par évaluation locale.

statut_candidature = le statut de candidature officiel, validé uniquement par un jury calibré qui mélange IA et humains, testés eux-mêmes pour les biais et la cohérence.

Cette distinction simple ferme la boucle auto-référentielle : aucun système ne peut se déclarer candidat AGI ou ASI simplement parce qu'il a obtenu de bons scores sur les benchmarks existants. La légitimité doit être accordée de l'extérieur.