Les benchmarks publics ont perdu le droit de décider seuls

De l'évaluation par les scores à l'architecture de validation

Jérôme Vétillard / Twingital-Institute – Avril 2026

Résumé

Les benchmarks IA publics tels que MMLU, HumanEval ou LMSYS Arena ont longtemps servi de signaux de compétence pour comparer des « foundation models » et orienter leur présélection. Cette fonction comparative persiste. En revanche, leur usage comme base suffisante de décision de déploiement en contexte régulé devient méthodologiquement fragile, pour deux ordres de raisons.

Le premier est temporel : l'érosion de l'indépendance effective des jeux de test par contamination, exposition prolongée ou adaptation aux plateformes d'évaluation ; la baisse du pouvoir discriminant à mesure que les modèles « frontier » se rapprochent du plafond ; et la divergence croissante entre performance mesurée et comportement observé en contexte opérationnel réel.

Le second est structurel : les benchmarks publics dominants sont conçus pour évaluer des modèles de type LLM ou des variantes proches. Les systèmes déployés en contexte régulé qui reposent souvent sur des architectures hétérogènes combinant modèles génératifs structurés, modèles tabulaires et composants spécialisés se trouvent en dehors du périmètre évalué. Ce qui est mesuré devient ce qui compte. Ce qui n'est pas mesuré devient structurellement invisible.

La distinction directrice proposée est celle entre *exogénéité apparente* et *exogénéité structurelle* : les benchmarks publics conservent souvent l'apparence d'un regard externe, alors même qu'ils sont progressivement réabsorbés dans la boucle d'optimisation des producteurs. La réponse n'est ni le rejet des benchmarks, ni leur fétichisation, mais la restauration d'une indépendance de validation par architecture. Pour les organisations opérant en contexte régulé, la sélection et la validation d'un système d'IA doivent être fondées sur une capacité interne de validation, documentée, reproductible, contextualisée au risque métier, et séparée structurellement de la fonction d'optimisation.

1. Introduction

La question de l'évaluation des modèles d'IA n'est pas un sujet technique périphérique. C'est un problème de gouvernance, et, dans les environnements régulés, un problème de preuve.

Le paradigme d'évaluation qui s'est imposé avec la montée des grands modèles de langage reposait sur une hypothèse implicite : un benchmark public correctement administré constitue un proxy suffisamment robuste de compétence générale pour guider la sélection d'un modèle. Dans cette configuration, la chaîne décisionnelle était simple. Les benchmarks publics fournissaient un signal de premier rang ; les décisions de présélection, et parfois de mise en production, s'alignaient sur ce signal.

Cette chaîne n'est plus tenable dans les mêmes termes. Non parce que les benchmarks seraient devenus inutiles, mais parce que les conditions de leur validité décisionnelle se sont détériorées dans le temps, par érosion de leur indépendance et de leur pouvoir discriminant ; et dès l'origine, par réduction de leur périmètre à une famille architecturale unique. Le diagnostic est désormais clair : l'évaluation évolue moins vite que les systèmes qu'elle prétend discriminer, et ne couvre qu'une fraction de ceux qu'elle prétend mesurer.

La thèse de cette note est la suivante : pour les organisations opérant en contexte régulé (santé, finance, pharma, assurance, conformité, fonctions juridiques critiques…), les benchmarks publics doivent être repositionnés comme signaux secondaires de veille et de présélection. La décision de déploiement doit reposer sur une validation interne contextualisée, documentée et reproductible, structurée comme une fonction distincte de la fonction d'optimisation du modèle.

Cette thèse s'inscrit dans un champ plus large, développé dans des travaux antérieurs sur la gouvernance agentique et l'architecture composite : la fiabilité d'un système d'IA ne se réduit jamais à la performance de son composant central. Elle se joue dans l'architecture qui l'entoure. Ce qui est vrai pour la gouvernance d'un agent, le composant ne gouverne pas le système, l'est tout autant pour l'évaluation d'un modèle : le score ne gouverne pas la décision.

2. Clarifications conceptuelles

Cinq distinctions doivent être stabilisées avant de poursuivre.

1. La première oppose *contamination* et *mémorisation*. La contamination désigne une atteinte au protocole d'évaluation : les données du benchmark, ou des versions très proches, réintègrent l'environnement d'entraînement ou de post-entraînement. La mémorisation est une propriété du modèle qui peut en résulter : le système restitue des motifs appris sans que le score reflète une compétence

réellement généralisable. Un benchmark peut être contaminé sans que toute performance soit régurgitation ; inversement, une bonne performance apparente peut masquer une mémorisation partielle difficile à détecter. Cela est arrivé lors de l'entrainement initial de « ToxTwin » et a conduit à un AUC très élevé du fait de la contamination du jeu d'entrainement.

2. La deuxième oppose *benchmark statique* et *benchmark dynamique*. Un benchmark statique présente l'avantage historique de la reproductibilité. Son défaut est qu'il se dégrade dès lors qu'il devient public, commenté, repris dans des jeux d'instructions ou exploité comme cible d'optimisation. Les benchmarks dynamiques tels que LiveBench [8] cherchent à réduire ce problème par renouvellement périodique des items. Ils ne suppriment pas tous les biais. Ils déplacent la frontière de validité.
3. La troisième oppose *hiérarchisation ordinale* et *validité absolue*. Un leaderboard peut conserver une valeur de tri comparatif même si ses scores absolus cessent d'être interprétables comme des garanties de robustesse en production. Le problème n'est pas que tout classement soit devenu nul. Le problème est que la conversion d'un écart de score public en hypothèse de fiabilité opérationnelle devient de moins en moins justifiable.
4. La quatrième oppose *érosion temporelle* et *biais de périmètre*. Un benchmark peut se dégrader dans le temps par contamination, saturation ou adaptation. Mais il peut aussi être structurellement inadéquat dès l'origine parce que son périmètre de mesure ne couvre pas l'architecture évaluée. Les deux problèmes sont distincts. Le premier est un problème de validité qui se détériore. Le second est un problème de validité qui n'a jamais existé.
5. La cinquième distinction est centrale. Elle oppose ***exogénéité apparente*** et ***exogénéité structurelle***. Un benchmark public donne l'apparence d'un regard externe : il semble provenir d'un dehors méthodologique indépendant du producteur. Mais cette extériorité peut n'être qu'apparente dès lors que le benchmark est connu, intégré, optimisé, anticipé ou contourné par les acteurs qu'il est censé mesurer. ***L'exogénéité structurelle suppose une indépendance effective entre la fonction qui optimise un système et la fonction qui l'évalue***. Le cœur de cette note est là : le problème des benchmarks publics n'est pas seulement leur imperfection, c'est l'érosion de cette indépendance effective. Un benchmark absorbé dans la boucle d'optimisation ne mesure plus. Il reflète.

3. Diagnostic : l'érosion de la validité prédictive

La contre-thèse est raisonnable : malgré leurs défauts documentés, les benchmarks publics conservent une valeur ordinale pour le screening, et les progrès méthodologiques pourraient suffire à préserver une partie de leur utilité. Ce n'est pas le point. Le point est que cette valeur ne suffit plus à autoriser un déploiement.

Trois mécanismes convergent pour expliquer cette érosion.

1. Le premier est la *contamination et les défauts internes*. L'existence même du draft NIST AI 800-2 [3] est révélatrice : si les benchmarks pouvaient encore être traités comme des instruments allant de soi, un tel niveau de précision méthodologique ne serait pas nécessaire. S'y ajoute le cas SWE-bench Verified, dont OpenAI a publiquement contesté la validité comme mesure des capacités « frontier » en codage : contamination croissante, nombre important de tests défectueux [10]. Lorsque le producteur d'un modèle « frontier » juge lui-même le benchmark inutilisable, le signal est difficile à ignorer.
2. Le deuxième est la *saturation*. À mesure que les meilleurs modèles se rapprochent du plafond sur certaines évaluations, quelques points d'écart deviennent plus difficiles à interpréter comme des différences opérationnellement significatives. Le Stanford AI Index 2026 documente cette convergence au sommet sur plusieurs benchmarks majeurs [1, 4]. Le benchmark continue d'ordonner, mais ordonne moins solidement qu'auparavant.
3. Le troisième est l'*écart entre benchmark et usage réel*. L'étude de Ribeiro et Lundberg (2022) sur les échecs de modèles NLP en contexte opérationnel [12], les analyses de Liang et al. sur les divergences entre dimensions HELM [11], et les retours terrain sur les agents de codage montrent un schéma récurrent : un modèle peut exceller sur un benchmark standardisé et échouer de manière inattendue sur des tâches multi-étapes, interactives ou fortement dépendantes d'un domaine métier. La relation entre score public et fiabilité de déploiement est trop instable pour servir seule de base à une décision à hauts enjeux.

Le problème n'est donc pas que les benchmarks mesurent imparfaitement. C'est qu'ils mesurent un mélange instable de compétence réelle, d'adaptation au test, de contamination et d'optimisation stratégique. Le score est un signal. La décision est une architecture.

4. Le biais de périmètre

Les trois mécanismes précédents décrivent une érosion dans le temps. Mais il existe un problème antérieur, qui ne relève pas de la dégradation : les benchmarks publics dominants n'évaluent qu'une famille architecturale.

MMLU, HumanEval, LMSYS Arena sont conçus pour mesurer des modèles de type LLM ou des variantes proches — raisonnement textuel, génération, programmation. Ce périmètre n'est pas un choix technique neutre. Il reflète une dynamique industrielle : la compétition entre laboratoires s'est structurée autour des modèles foundation généralistes, et les benchmarks se sont alignés sur cette course. Ce qui est mesuré devient ce qui compte. Ce qui n'est pas mesuré devient structurellement invisible.

Or les systèmes déployés en contexte régulé reposent rarement sur un LLM seul. Ils mobilisent ***des architectures hétérogènes***, des modèles génératifs structurés (CT-GAN, TVAE), modèles tabulaires (Random Forest, CatBoost, XGBoost), des composants spécialisés par domaine, ***dont les propriétés pertinentes*** (robustesse statistique, stabilité en extrapolation, cohérence des distributions générées, comportement sur des cohortes à faible effectif) ***ne sont capturées par aucun leaderboard public***.

Le cas est concret. Dans le cadre de TweenMe®, notre plateforme de génération de jumeaux numériques mobilisant plus de 25 modèles spécialisés, les benchmarks publics n'évaluent qu'une fraction marginale du système. La validation repose nécessairement sur des protocoles internes, alignés sur les cas d'usage, les distributions de données métier et les contraintes opérationnelles. Ces protocoles sont plus exigeants sur certaines dimensions que les benchmarks publics (robustesse, stabilité, cohérence statistique, comportement en extrapolation). Ils sont aussi moins lisibles, moins comparables, et moins mobilisables comme signal externe de crédibilité.

La tension est structurante. Les benchmarks publics offrent un langage commun de positionnement au prix d'une réduction du périmètre évalué. Les benchmarks internes offrent une pertinence opérationnelle au prix d'une moindre reconnaissance. Cette tension ne se résout pas par le choix d'un camp. Elle confirme la nécessité de l'architecture à deux niveaux décrite plus loin : benchmarks publics comme langage commun, validation interne comme base de décision.

Surtout, ce biais de périmètre rend caduque l'objection spontanée selon laquelle les benchmarks « vont s'améliorer ». Même un benchmark public renouvelé, décontaminé, méthodologiquement irréprochable, reste hors sujet pour un système qui ne repose pas principalement sur un LLM. Un benchmark LLM-centrique amélioré ne cesse pas d'être LLM-centrique. L'amélioration du protocole ne corrige pas l'inadéquation du Benchmark à évaluer une architecture Deep Neural Network différente (non-LLM).

5. De Goodhart au problème de gouvernance

Dès lors qu'une mesure publique acquiert une forte valeur réputationnelle, commerciale ou financière, elle devient une cible d'optimisation. Lorsque les leaderboards influencent la valorisation, le financement, le recrutement ou l'adoption client, il devient rationnel pour les laboratoires d'optimiser non seulement leurs modèles, mais aussi la manière dont leurs modèles se présentent dans les dispositifs d'évaluation.

L'intuition est celle formulée par Goodhart en 1975 [9], une régularité statistique observée tend à se dégrader lorsqu'on exerce sur elle une pression de contrôle, puis reformulée plus largement par Strathern (1997) [13] dans le contexte de l'audit. La mesure qui réussit trop bien finit par mesurer sa propre influence.

Ce raisonnement ne doit pas être transformé en accusation indiscriminée. Il n'est pas nécessaire de supposer de mauvaises pratiques systématiques. Il suffit de reconnaître qu'à partir du moment où un benchmark public devient un actif de signalement, il existe une incitation structurelle à l'optimiser ou à s'y adapter. Cette incitation fragilise son statut de mesure indépendante pour des décisions à fort enjeu.

Le point décisif est architectural, pas moral. Même en supposant des producteurs de bonne foi, un système d'évaluation public, connu et valorisé finit par être absorbé dans la boucle d'optimisation du système évalué. Il perd alors sa fonction d'extérieur crédible. L'exogénéité cesse d'être structurelle. Elle ne subsiste que comme apparence.

Ce mécanisme se combine avec le biais de périmètre pour produire un double angle mort : les benchmarks publics mesurent de moins en moins bien ce qu'ils prétendent mesurer (érosion temporelle), et ne mesurent pas du tout ce qu'ils ne prétendent pas mesurer (biais de périmètre). La conjonction des deux rend leur statut décisionnel insoutenable pour les systèmes hétérogènes en contexte régulé.

6. Proposition : l'architecture de validation en trois couches

La conséquence n'est pas l'abolition des benchmarks. C'est leur repositionnement dans une architecture plus large où chaque couche a un rôle distinct, une portée définie et un niveau d'autorité explicite.

- *Première couche : veille et présélection.* Les benchmarks publics y conservent une utilité réelle. Ils permettent de suivre l'état de l'art, d'identifier des familles de modèles, d'observer des dynamiques de progrès et de constituer un premier ensemble de candidats. LiveBench [8] montre qu'il est possible de repousser la frontière de contamination par renouvellement périodique. HELM [11] montre qu'une évaluation utile peut être multidimensionnelle plutôt que compressée dans un unique score commode mais trompeur. Ce qui importe n'est pas l'adoption d'un outil unique, mais le déplacement vers des évaluations plus difficiles à contaminer et plus riches analytiquement. Cette couche oriente. Elle ne décide pas.
- *Deuxième couche : validation interne contextualisée.* C'est ici que la décision se joue. Les jeux d'évaluation doivent être alignés sur le domaine, le niveau de risque, les flux de travail réels, la distribution des cas faciles et difficiles, les modes d'échec attendus, les exigences de robustesse et de reproductibilité. Cette couche doit documenter ce qu'elle mesure, ce qu'elle ne mesure pas, sa stabilité et son domaine de validité.

Concrètement, cela implique la construction de jeux de test internes dérivés de données métier réelles ou réalistes, incluant des cas adversariaux spécifiques au domaine, des scénarios multi-étapes reproduisant les chaînes d'action en production, et des métriques alignées sur le risque métieret non pas sur la précision générique. Les protocoles de validation doivent être documentés avec le même niveau de rigueur qu'un protocole d'essai clinique ou un dossier de validation réglementaire : version du modèle, version du jeu de test, conditions d'exécution, seuils d'acceptation, critères d'exclusion.

Pour les systèmes multi-architectures (ceux qui combinent LLM, modèles tabulaires, modèles génératifs structurés et composants spécialisés) cette couche devient le seul lieu où l'évaluation est pertinente. Les benchmarks publics, par construction LLM-centriques, ne couvrent ni la robustesse des distributions synthétiques, ni la stabilité des modèles tabulaires, ni la cohérence des pipelines d'orchestration. La validation interne n'est alors pas un complément. C'est le signal primaire.

L'intérêt du draft NIST AI 800-2 [3] n'est pas d'imposer déjà un standard contraignant, mais de rendre visible la direction : l'évaluation doit devenir méthodologiquement explicite, statistiquement documentée et interprétée avec prudence. Le mouvement est engagé. Il ne sera pas réversible.

- *Troisième couche : séparation organisationnelle*. La fonction qui tune, sélectionne, optimise ou intègre le modèle ne doit pas être seule juge de la validité qui autorise son usage. Cette séparation n'instaure pas une exogénéité institutionnelle parfaite. Elle restaure une *exogénéité de design* suffisante pour réduire les conflits de rôle, documenter les arbitrages et rendre l'autorisation de déploiement opposable.

La logique est celle de tout système de contrôle sérieux : celui qui prescrit ne se valide pas lui-même. En finance, c'est le principe de ségrégation des fonctions. En essais cliniques, c'est la séparation entre investigateur et comité d'évaluation. En déploiement de modèles d'IA en contexte régulé, c'est le même raisonnement et il n'est pas encore intégré par la majorité des organisations.

7. Contribution : exogénéité apparente, exogénéité structurelle

La distinction entre exogénéité apparente et exogénéité structurelle est proposée ici comme cadre analytique et non pas comme catégorie réglementaire formalisée. Sa fonction est explicative et architecturale.

L'exogénéité apparente est celle d'un benchmark public qui semble provenir d'un dehors méthodologique indépendant, mais dont l'indépendance effective s'est érodée par exposition, optimisation, contamination ou absorption dans les pratiques

d'entraînement. L'exogénéité structurelle suppose une séparation effective (par design, par organisation, par protocole) entre la fonction qui optimise et la fonction qui évalue. Elle ne requiert pas l'indépendance parfaite. Elle requiert une indépendance architecturée, documentée, opposable.

Ce cadre a une portée qui dépasse l'évaluation des modèles d'IA. Il décrit un problème récurrent dans tout système d'évaluation à forts enjeux : l'audit financier absorbé par le conseil, la notation de crédit absorbée par la structuration, l'évaluation clinique absorbée par le promoteur. Chaque fois, le mécanisme est le même : la mesure qui acquiert une valeur de signalement est progressivement réabsorbée dans la boucle d'optimisation de ceux qu'elle est censée mesurer. La réponse, dans chaque cas, a été architecturale : séparation des fonctions, rotation, indépendance par design.

L'évaluation des foundation models entre dans cette catégorie. Et la réponse sera la même.

8. Articulation : de la gouvernance des modèles à la gouvernance des systèmes

Cette analyse prolonge un raisonnement développé dans deux travaux antérieurs.

Dans « La gouvernance agentique ne viendra pas des modèles », la thèse était que la gouvernabilité d'un système agentique ne dépend pas de la qualité du modèle mais de l'architecture dans laquelle il opère. Le triplet action-space / autonomie / réversibilité structure le régime de délégation, indépendamment de la performance du composant.

Le parallèle est direct. Ce qui est vrai pour la gouvernance d'un agent l'est aussi pour l'évaluation d'un modèle : la fiabilité ne se lit pas dans le score du composant. Elle se lit dans l'architecture de validation qui entoure la décision de déploiement.

Dans « Au-delà du paradigme LLM-centré : architecture agentique composite pour les jumeaux numériques en environnement régulé », la thèse complémentaire était que le modèle de langage n'est ni aboli, ni absolutisé. Il est *repositionné* comme composant parmi d'autres dans une architecture qui le dépasse. La même opération s'applique aux benchmarks : ils ne sont ni rejetés, ni fétichisés. Ils sont repositionnés dans une architecture de validation qui les dépasse.

Le benchmark n'est ni aboli, ni absolutisé : Il est repositionné. C'est le même mouvement intellectuel, appliqué à l'évaluation.

Le biais de périmètre identifié en section 4 renforce cette articulation. L'architecture composite repose, par design, sur des modèles qui ne sont pas des LLM : modèles tabulaires, génératifs structurés, composants d'orchestration. Les benchmarks publics LLM-centriques ne peuvent pas évaluer ces systèmes, non par imperfection, mais par

construction. La validation interne contextualisée n'est pas un luxe méthodologique pour ces architectures. C'est la seule option qui existe.

9. Insuffisance du paradigme de marché

Les solutions actuellement proposées par le marché (guardrails, monitoring post-déploiement, RLHF, red teaming) adressent le problème au niveau du composant. Elles améliorent le modèle ou surveillent ses sorties. Elles ne construisent pas une architecture de validation.

Un guardrail empêche un modèle de produire certaines sorties. Il ne garantit pas que le modèle est adapté au domaine métier. Le monitoring observe des métriques de performance en production. Il ne remplace pas un protocole de validation *avant* déploiement. Le red teaming teste la robustesse du modèle face à des attaques. Il ne traite pas la question de la validité décisionnelle du score qui a justifié le déploiement.

Ces outils ne sont pas mauvais. Ils opèrent au mauvais niveau. La gouvernance de l'évaluation est un problème de système. Les solutions de marché sont encore très majoritairement des solutions de composant. La même erreur de niveau que pour la gouvernance agentique et donc la même conséquence : tant qu'on traite un problème de système comme un problème de composant, on ne le résout pas.

10. Terrain d'implémentation : la validation dans PREDICARE et TweenMe

Le programme PREDICARE (médecine prédictive territoriale en zone de désertification médicale) et le cadre TweenMe® illustrent concrètement ce que signifie une architecture de validation séparée de la fonction d'optimisation. Ce ne sont pas des preuves générales du cadre proposé. Ce sont des terrains où le cadre est devenu pratique opérationnelle.

Dans le cadre d'un travail de validation d'une cohorte synthétique oncologique européenne (poster ISPOR 2025), la question de l'évaluation s'est posée dans des termes qui illustrent exactement la distinction entre exogénéité apparente et exogénéité structurelle. Les résultats de fidélité opérationnelle obtenus (métriques de classification et tests de survie) n'ont de valeur que parce que le protocole de validation a été structurellement séparé de la boucle de génération : la cohorte synthétique a été produite par un pipeline, la validation a été conduite sur la cohorte réelle, selon un protocole documenté, avec des métriques prédéfinies, des critères d'interprétation explicites, et une analyse des limites publiée.

Ce qui a été validé (la fidélité opérationnelle de la cohorte synthétique pour les tâches aval) a été explicitement distingué de ce qui n'a pas été validé : L'indistinguabilité

statistique générale. Le résultat n'est pas un score sur un leaderboard. C'est une preuve contextualisée, reproductible, limitée dans son domaine de validité, et structurellement indépendante de la fonction d'optimisation.

Ce cas illustre également le biais de périmètre. Le pipeline TweenMe mobilise des CT-GAN, des modèles de Fine & Gray, SurvTRACE, des forêts aléatoires… Aucun de ces composants n'est évaluable par un benchmark LLM-centrique. La validation interne n'est pas ici un complément de luxe. C'est la seule validation qui existe pour cette classe de systèmes.

De même, dans PREDICARE, le choix des modèles embarqués dans le jumeau numérique patient n'a pas été fondé sur un benchmark public. Il a été fondé sur une validation interne contextualisée : jeux de test dérivés des données du protocole clinique, métriques alignées sur le risque clinique (sensibilité sur les alertes, spécificité sur les faux positifs qui génèrent de la fatigue d'alerte), seuils d'acceptation définis par le comité médical et non pas par l'équipe de développement. La séparation des fonctions n'est pas un principe abstrait. C'est une architecture opérationnelle.

11. Ce que cela change pour un CTO et pour un COMEX

Pour un CTO, la conséquence est une inversion de hiérarchie décisionnelle. Le benchmark public cesse d'être le support principal de décision. Il devient un instrument d'orientation de premier tri. La décision migre vers une fonction d'évaluation interne gouvernée. Cela implique de constituer une capacité d'évaluation (pas nécessairement un laboratoire), mais une fonction capable de concevoir des jeux de test contextualisés, de documenter les protocoles, d'analyser les échecs et de maintenir un référentiel de validation distinct des supports marketing des fournisseurs.

Pour les organisations qui opèrent des systèmes multi-architectures, cette capacité n'est pas optionnelle. Elle est la condition d'existence même d'une évaluation pertinente puisque les benchmarks publics ne couvrent pas le périmètre du système déployé.

Pour un COMEX, le sujet n'est pas technique au sens étroit. Il relève du contrôle interne et de la qualité des preuves mobilisées pour autoriser un système à entrer dans un processus métier. La question n'est plus : quel modèle a le meilleur score public ? Elle devient : sur quelle base méthodologique opposable avons-nous estimé que ce modèle pouvait être déployé dans ce contexte précis ?

Quatre implications concrètes en découlent.

1. Premièrement, l'entreprise doit formaliser une *politique de preuve* : quel niveau de démonstration est requis selon la criticité d'usage ? Quelle combinaison de benchmarks publics, d'évaluations internes, de tests adversariaux et de supervision humaine conditionne l'autorisation de déploiement ?

2. Deuxièmement, elle doit *séparer les rôles* : l'équipe qui pousse l'intégration d'un modèle ne contrôle pas seule la conclusion de validité.
3. Troisièmement, elle doit *documenter les arbitrages* de manière opposable.
4. Quatrièmement, elle doit accepter que la gouvernance de l'évaluation devienne une question d'architecture de décision et non pas de procurement ou de benchmark shopping.

L'intérêt de cette approche est qu'elle s'aligne avec le sens des évolutions institutionnelles. Le NIST pousse vers des pratiques documentées d'évaluation [3, 6]. L'AI Office européen développe des outils et méthodologies pour évaluer les GPAI models [7]. Ces cadres demeurent en cours de stabilisation. Mais la direction est claire : les exigences de preuve, de méthode et d'opposabilité vont se durcir. Les organisations qui attendent un standard finalisé pour agir se trouveront en retard sur un mouvement déjà engagé.

12. Discussion et limites

Ce cadre présente des limites qu'il convient d'expliciter.

1. Première limite : la hiérarchisation ordinale conserve une valeur. Il serait excessif de prétendre qu'un leaderboard public ne dit plus rien. Il continue à fournir des signaux utiles de présélection et de veille. La thèse n'est pas que les benchmarks sont morts. Elle est qu'ils ne peuvent plus décider seuls.
2. Deuxième limite : toutes les dégradations ne sont pas uniformes. Tous les benchmarks ne se dégradent pas au même rythme, tous les domaines ne sont pas exposés de la même manière, et tous les usages n'exigent pas le même niveau de preuve. Le cadre proposé vise les décisions haute-stakes en contexte régulé. Il ne prétend pas que le screening d'un outil conversationnel grand public exige la même rigueur.
3. Troisième limite : la notion d'exogénéité structurelle est proposée comme cadre analytique. Elle n'est pas une catégorie réglementaire formalisée comme telle. Sa fécondité est explicative et architecturale car elle permet de nommer un problème que tout le monde reconnaît sans le formaliser.
4. Quatrième limite : le coût de la validation interne n'est pas traité. Construire une capacité d'évaluation interne contextualisée a un coût (en compétences à développer, en données, en temps, en infrastructure). Ce coût n'est pas négligeable, et il peut constituer une barrière pour des organisations de taille moyenne. Le cadre identifie ce qui doit être fait. Il ne prétend pas que ce soit gratuit.
5. Cinquième limite : le biais de périmètre décrit en section 4, s'il est structurellement incontestable, ne concerne pas de manière identique toutes les organisations. Celles qui déploient exclusivement des LLM généralistes restent

dans le périmètre des benchmarks publics même si les problèmes d'érosion temporelle demeurent. L'argument du biais de périmètre prend toute sa force pour les systèmes multi-architectures en contexte régulé. Il ne doit pas être sur-généralisé.

6. Sixième limite : les cadres institutionnels de 2026 sont en cours de stabilisation. Le NIST AI 800-2 est un draft de bonnes pratiques volontaires. L'AI Office européen monte en régime. Il convient de parler d'orientation forte plutôt que de standard achevé.

13. Conclusion

Les benchmarks publics n'ont pas cessé d'être utiles. Ils ont cessé d'être suffisants.

Leur faiblesse tient à deux ordres de raisons. Le premier est temporel : contamination, saturation, adaptation au test érodent progressivement leur validité prédictive. Le second est structurel : conçus pour une famille architecturale unique, ils ne couvrent pas les systèmes hétérogènes qui composent la réalité des déploiements en contexte régulé. Un benchmark LLM-centrique amélioré ne cesse pas d'être LLM-centrique.

Dans les deux cas, le mécanisme profond est le même : l'exogénéité devient apparente. Les benchmarks perdent la distance méthodologique qui fondait leur valeur décisionnelle, soit parce qu'ils sont absorbés dans la boucle d'optimisation, soit parce qu'ils ne mesurent pas ce qui est effectivement déployé.

Le problème est de même nature que celui de la gouvernance agentique : tant qu'on cherche la fiabilité dans le composant, on ne la trouve pas, parce qu'elle se joue dans l'architecture. Un modèle excellent évalué par un benchmark absorbé dans sa propre boucle d'optimisation n'est pas un modèle validé. C'est un modèle bien classé ce qui n'est pas la même chose.

Pour un CTO, cela impose de construire une fonction de validation interne. Pour un COMEX, cela impose de traiter l'évaluation comme un sujet de contrôle de décision, pas d'achat technologique. Pour l'organisation, cela impose une architecture de validation fondée sur trois principes : benchmarks publics comme signaux secondaires de veille, évaluation interne contextualisée comme signal primaire, et séparation structurelle entre optimisation et autorisation de déploiement.

Les benchmarks publics ont perdu le droit de décider seuls. Ce n'est pas une critique. C'est un diagnostic et une indication architecturale claire.

Notes

[1] Stanford Human-Centered Artificial Intelligence. *AI Index Report 2026*. Stanford University, 2026.

[2] Kontorovich, Aryeh, et al. "Benchmarking Large Language Models Under Data Contamination: A Survey from Static to Dynamic Evaluation." *arXiv*, 2025.

[3] National Institute of Standards and Technology. *AI 800-2: Automated Evaluation of AI Systems: Practices and Considerations*. Initial Public Draft, janvier 2026.

[4] Stanford Human-Centered Artificial Intelligence. *AI Index Report 2026*, section "Technical Performance".

[5] La distinction entre exogénéité apparente et exogénéité structurelle est proposée ici comme cadre analytique pour décrire la perte d'indépendance effective des benchmarks publics et la nécessité d'une validation séparée par design.

[6] National Institute of Standards and Technology. *AI 800-2*, sections consacrées à la documentation, au benchmark design, à l'analyse statistique et à l'interprétation des résultats.

[7] European Commission, AI Office. Documentation institutionnelle relative aux GPAI models, méthodes d'évaluation et montée en régime de l'application de l'AI Act, 2025-2026.

[8] White, L., Vinitsky, E., and Sridhar, N. "LiveBench: A Contamination-Limited Benchmark for Large Language Models." *arXiv preprint*, 2024.

[9] Goodhart, Charles A. E. "Problems of Monetary Management: The U.K. Experience." Reserve Bank of Australia, 1975.

[10] OpenAI. "Why SWE-bench Verified No Longer Measures Frontier Coding Capabilities." 2026.

[11] Liang, Percy P., Bommasani, Rishi, Lee, Tony, et al. "Holistic Evaluation of Language Models." *arXiv preprint*, 2023.

[12] Ribeiro, Marco Tulio, and Lundberg, Scott. "Adaptive Testing and Debugging of NLP Models." *Proceedings of the 60th Annual Meeting of the ACL*, 2022.

[13] Strathern, Marilyn. "'Improving Ratings': Audit in the British University System." *European Review*, vol. 5, no. 3, 1997, pp. 305-321.

Bibliographie

Commission européenne. *Regulation (EU) 2024/1689 on artificial intelligence (AI Act)*. Official Journal of the European Union, 2024.

Goodhart, Charles A. E. "Problems of Monetary Management: The U.K. Experience." Reserve Bank of Australia, 1975.

Hendrycks, Dan, Burns, Collin, Basart, Steven, et al. "Measuring Massive Multitask Language Understanding." *arXiv preprint*, 2020.

Kontorovich, Aryeh, et al. "Benchmarking Large Language Models Under Data Contamination: A Survey from Static to Dynamic Evaluation." *arXiv*, 2025.

Liang, Percy P., Bommasani, Rishi, Lee, Tony, et al. "Holistic Evaluation of Language Models." *arXiv preprint*, 2023.

National Institute of Standards and Technology. *AI 800-2: Automated Evaluation of AI Systems: Practices and Considerations*. Initial Public Draft, 2026.

OpenAI. "Why SWE-bench Verified No Longer Measures Frontier Coding Capabilities." 2026.

Ribeiro, Marco Tulio, and Lundberg, Scott. "Adaptive Testing and Debugging of NLP Models." *Proceedings of the 60th Annual Meeting of the ACL*, 2022.

Stanford Human-Centered Artificial Intelligence. *AI Index Report 2026*. Stanford University, 2026.

Strathern, Marilyn. "'Improving Ratings': Audit in the British University System." *European Review*, vol. 5, no. 3, 1997, pp. 305-321.

White, L., Vinitsky, E., and Sridhar, N. "LiveBench: A Contamination-Limited Benchmark for Large Language Models." *arXiv preprint*, 2024.