Performance sur benchmark n'est pas déployabilité : trois ports de fiabilité, pas trois métriques

Trois ruptures structurelles séparent l'évaluation benchmark du régime de production : split, calibration, domaine d'applicabilité. La déployabilité s'installe par architecture, pas par score.

Le problème

Un mois sur deux, un nouveau modèle bat l'état de l'art sur un benchmark public. Le communiqué circule. Les équipes produits font remonter la nouvelle au COMEX. Dans la moitié des cas, le modèle finit en preuve de concept gelée, ou en déploiement silencieusement défaillant. Ce n'est pas un problème d'exécution. C'est un problème de mesure : l'industrie continue de traiter le score sur un benchmark comme la preuve qu'un modèle tient en production, alors que les deux régimes d'évaluation n'ont presque rien en commun.

L'écart n'est pas un défaut récent. Il est structurel et il s'aggrave. Les benchmarks sont devenus plus saturés ; les marges entre modèles se réduisent ; les leaderboards sont d'autant plus mobilisés comme outils de communication qu'ils discriminent moins. Pendant ce temps, la production reçoit des requêtes que le benchmark n'a jamais vues, sous des distributions que le hold-out (la fraction réservée à l'évaluation) n'a jamais simulées.

Pourquoi les solutions actuelles échouent

La position dominante est simple : on évalue, on déploie, on surveille. On ajoute du monitoring si nécessaire. Ce raisonnement repose sur une hypothèse rarement explicitée : que le score sur le benchmark constitue une mesure utile de la performance opérationnelle, modulo un facteur de dégradation absorbé par le monitoring aval.

Cette hypothèse échoue sur trois ruptures, déjà nommées dans les posts de cette semaine.

1. La première rupture est celle du ***split.*** Le benchmark utilise généralement un partage aléatoire. La production reçoit ses données dans un ordre temporel, avec dérive de population et de pratiques. Un modèle qui a appris une fuite temporelle n'apparaît jamais comme tel sur un split aléatoire, il apparaît comme excellent. C'est exactement l'objet du post de vendredi : si la performance s'effondre quand on passe au split temporel, le modèle apprenait la fuite, pas la tâche. Un split aléatoire en industrie de santé est un test de mémoire ; un split temporel est un test de généralisation.

2. La deuxième rupture est celle de la ***calibration***. Le score d'AUC (aire sous la courbe ROC) mesure un ordre. Il peut rester très bon alors que les probabilités produites par le modèle sont systématiquement biaisées. En clinique, en pharmacovigilance, en triage opérationnel, on n'exploite pas un ordre, on exploite des probabilités, dont les seuils déclenchent des actions coûteuses. Une AUC de 0,90 sans calibration n'est pas un modèle utilisable ; c'est un modèle classant.

3. La troisième rupture est celle du ***domaine d'applicabilité (AD).*** Le hold-out couvre la distribution de l'entraînement. La production reçoit des requêtes hors zone (un patient avec une mutation rare, une molécule structurellement éloignée, une comorbidité absente du jeu d'entraînement). Sans AD signé, le modèle répond. Il extrapole, sans signal qu'il extrapole.

Le modèle alternatif

L'alternative ne consiste pas à abandonner les benchmarks. Elle consiste à cesser de les confondre avec une mesure de déployabilité. Trois ports de fiabilité doivent être installés, et leur installation est une décision d'architecture autant que d'évaluation.

1. Premier port : split discipliné. Temporel pour les données qui dérivent dans le temps (épidémiologie, pharmacovigilance, signaux cliniques). Par scaffold pour les données moléculaires, où la similarité structurelle crée une fuite invisible au split aléatoire. Le coût d'un split correct est une chute apparente de score. C'est exactement le coût qu'on cherche à payer en amont, plutôt qu'en production.

2. Deuxième port : calibration explicite. Méthode isotonique ou équivalent, sur un set indépendant du test, avec diagramme de fiabilité publié. Ce port transforme un score d'ordre en probabilité utilisable.

3. Troisième port : AD signé, vérifiable à l'inférence. Une requête hors domaine reçoit un rejet typé, pas une prédiction silencieuse. Le post de lundi a posé la formule : «un score sans split temporel, sans calibration, sans AD, n'est pas une mesure de fiabilité, c'est un colifichet.»

Sur PREDICARE, plateforme de prédiction pharmacologique, les trois ports ne sont pas des étapes de validation amont, ils sont la structure du pipeline. Sur ToxTwin, le jumeau toxicologique, le split par scaffold est imposé en amont de tout entraînement, et l'AD est versionné avec le modèle. Ces instances ne prouvent pas que la doctrine est universellement bonne. Elles montrent qu'elle est implémentable, et que ses coûts sont mesurables.

Implication CTO / COMEX

La conséquence décisionnelle est nette. Un programme d'IA dont le budget de validation est inférieur au budget de modélisation est un programme qui achète des trophées et finance ensuite l'incident en production. Le benchmark public est une condition nécessaire de la déployabilité ; il n'en est jamais la preuve. Le post de mercredi a traité l'objection de la communauté ML : si elle durcit ses propres protocoles (splits temporels, splits par scaffold, décalages de distribution explicites) c'est qu'elle reconnaît implicitement que le score brut ne mesurait pas ce que le marché lui faisait dire.

Une question simple sépare les deux régimes en COMEX : « combien votre meilleur modèle perd-il, en performance, quand vous ré-évaluez sur la dernière fenêtre temporelle non vue, après calibration, restreint à l'AD ? » Si le chiffre n'existe pas, le modèle n'est pas déployable, il est éligible à un test. Si le chiffre existe et qu'il est petit, vous gouvernez votre IA. S'il est grand, vous savez ce qu'il faut financer. Le reste (communiqué, leaderboard, slide) est de la communication, pas de la fiabilité.

[Série : Benchmark ≠ Production / article de clôture]