GUIDE · GRATUIT

Comment savoir si votre IA fait vraiment ce que vous pensez

Presque toutes les organisations qui déploient de l'IA valident au ressenti, sur quelques exemples qui marchent. Ce guide explique ce qu'est une évaluation sérieuse, pourquoi elle ne doit pas être confiée à ceux qui ont construit le système, et ce qu'elle vous ouvre une fois qu'elle existe.

SynaLinks est un cabinet de recherche appliquée aux systèmes neuro-symboliques, basé à Toulouse. Nous construisons les jeux de données et les évaluations qui décident si un système d'IA tient en production, ainsi que les frameworks open source qui les produisent.

LEVER UNE CONFUSION D'ABORD

Trois fausses évaluations.

Les scores publics

Ils comparent des modèles entre eux sur des tâches génériques. Ils ne disent rien de ce que le vôtre fait sur vos dossiers, ou pour vos clients. Un modèle premier au classement peut être mauvais chez vous.

Les tests manuels, ou « vibe-testing »

Quelqu'un pose quinze questions, lit les réponses, trouve ça bien. Ni reproductible, ni comparable d'une version à l'autre. Et les quinze questions sont celles auxquelles le système sait répondre.

Le modèle qui se note lui-même, ou « LLM-as-a-judge »

Souvent la seule méthode praticable à grande échelle, mais un juge non validé est une opinion de plus. Étalonnez-le contre des jugements humains et mesurez son accord avant de lui faire confiance.

Des cas tirés du réel
vos dossiers, vos vraies questions, échecs compris

Des critères écrits
ce qu'est une bonne réponse, cas par cas

Un harnais
qui rejoue le tout à la demande

=

Une mesure
Reproductible. Comparable d'une version à l'autre.

Rejouer des cas est facile. Décider si une réponse est juste, beaucoup moins.

Ce verdict décide de tout, et la plupart des outils ne savent le rendre qu'en demandant son avis à un autre modèle.

C'est notre spécialité. Nous construisons des systèmes **neuro-symboliques** : un modèle de langage adossé à une couche symbolique, faite de code, de logique et de contraintes. C'est elle qui rend la vérification automatique. Une réponse n'est plus jugée, elle est contrôlée : une règle passe ou échoue, une requête ramène la bonne ligne ou non. Le verdict est identique à chaque exécution, il s'explique, et il se rejoue pour une fraction du coût d'un juge modèle.

LE POINT QUI FÂCHE

On ne note pas sa propre copie.

Le problème n'a rien à voir avec l'honnêteté : il est structurel, et il joue contre les gens les plus intègres du monde.

LE CONSTRUCTEUR S'ÉVALUE

Choisit l'architecture →

Choisit les cas de test

Une seule hypothèse sur ce qui compte

Le cas qu'il n'a pas anticipé en construisant, il ne l'anticipe pas non plus en testant.

UN TIERS ÉVALUE

Construit le système →

Demande ce que « juste » veut dire

Deux hypothèses indépendantes

Il ne peut pas lire l'intention dans le code : il est obligé de poser la question au métier.

Et l'incitation va dans le mauvais sens. Une évaluation qui conclut « ça ne marche pas » coûte à son auteur la phase suivante de la mission.

Rien de tout cela n'est propre à l'IA. On ne laisse pas une entreprise certifier ses propres comptes ni auditer sa propre sécurité.

Demandez qui a écrit l'évaluation de votre système. Si c'est celui qui a écrit le système, vous connaissez déjà sa conclusion.

CE QUE ÇA DÉBLOQUE

Six choses impossibles sans mesure.

On la présente comme une formalité de fin de projet, alors que tout le reste repose dessus.

SANS ÉVALUATION

- Le système répond
- Quelqu'un trouve ça bien
- ?

La chaîne s'arrête. Rien à optimiser, rien à comparer, rien à prouver.

AVEC ÉVALUATION

- Le système répond
- La suite mesure
- Les échecs désignent la cause
- Vous corrigez

↺ et on remesure

La boucle se ferme, et un agent peut la faire tourner à votre place.

01 **Changer de modèle quand vous le décidez**

Les fournisseurs déprécient leurs modèles à leur calendrier, pas au vôtre. Sans évaluation, migrer est un saut dans le vide. Avec, vous rejouez la suite sur le candidat et vous décidez sur des chiffres.

02 **Payer moins cher pour la même qualité**

Sans mesure, tout le monde reste sur le plus gros modèle par peur. Avec, vous essayez le petit modèle ouvert qui tourne chez vous et vous voyez ce que la qualité y perd. Souvent, rien.

03 **Automatiser l'amélioration au lieu d'en débattre**

Une évaluation est un signal. Sans lui, rien à optimiser. Avec lui, un agent cherche à votre place : prompts, exemples, distillation sur votre domaine. L'amélioration devient une boucle qui tourne.

04 **Faire de la recherche sur votre propre cas**

Lancez vingt variantes d'architecture, gardez celle qui gagne. Les décisions se règlent par des expériences, pas par la séniorité de celui qui parle.

05 **Livrer sans retenir votre souffle**

Chaque changement passe la suite avant la production : un prompt, un index, une source, une version de modèle. Les régressions se voient chez vous, pas chez vos clients.

06 **Répondre à « comment le savez-vous ? »**

Devant un conseil, un grand client ou un régulateur, la question est toujours la même. Une évaluation rejouable transforme une conviction en preuve. C'est ce que l'AI Act commence à exiger.

À QUOI ÇA RESSEMBLE

Cinq propriétés d'une bonne évaluation.

Tirée du réel

les cas viennent de vos dossiers et de vraies questions, échecs compris. Un jeu de test qui ne contient que des cas faciles mesure votre optimisme.

À critères explicites

« bonne réponse » est écrit cas par cas : la bonne source, le bon montant, le refus quand l'information manque. Un critère qui ne s'écrit pas ne se mesure pas.

Reproductible

relancée demain, elle donne le même résultat. Versionnée avec le code, exécutée automatiquement.

Discriminante

elle doit pouvoir échouer. Une évaluation que tout passe à 98 % vous rassure au lieu de vous apprendre quelque chose.

À vous

le jeu de données, les critères et le harnais restent chez vous, et se réutilisent à chaque nouveau modèle.

PAR OÙ COMMENCER

Deux étapes à faire sans nous.

1 **Collectez trente échecs**

Pas trente questions : trente cas où votre système a mal répondu, dans la vraie vie. Si personne ne les a notés, c'est déjà un résultat, et le premier problème à régler.

2 **Écrivez ce qu'aurait été la bonne réponse**

Cas par cas, avec la source qui le prouve. C'est fastidieux, c'est là qu'est le travail, et c'est ce qui révèle les désaccords internes sur ce que « juste » veut dire.

Arrêtez-vous là. Ces deux étapes ne coûtent que de l'attention et produisent la seule chose qui manque à tout le monde : des **exemples étiquetés**, tirés de votre métier, où quelqu'un a écrit ce que « juste » voulait dire.

C'est la matière première de la suite. Ils servent à calibrer un juge automatique et à vérifier qu'il est d'accord avec vous avant qu'on lui fasse confiance à grande échelle. C'est là que nous prenons le relais : couvrir la variété réelle des cas, valider le juge, industrialiser le harnais, et réparer les données que vos trente échecs auront mises en cause.