



CHECKLIST 8 :

ÉVALUATION, TEST ET SURVEILLANCE DES MODÈLES

Comment s'assurer qu'un modèle d'IA générative produit des résultats fiables, pertinents et de qualité ? L'évaluation objective des performances est un défi majeur qui nécessite une méthodologie rigoureuse combinant métriques quantitatives, tests humains et surveillance continue. Cette checklist vous fournit les outils et processus pour évaluer, tester et monitorer vos modèles tout au long de leur cycle de vie.

→ Cadre d'Évaluation

- Définir un cadre de test objectif pour passer d'une évaluation subjective à des résultats mesurables.
- Établir des rubriques d'évaluation (statiques ou adaptatives) avec des critères clairs pour juger la qualité des réponses.
- Utiliser une combinaison de méthodes : évaluation humaine, métriques calculées, et évaluation par un autre LLM (*LLM-as-a-judge*).

→ Tests de Performance

- Utiliser une liste de questions types ("*pain test questions*") représentatives des cas d'usage pour comparer les performances de différents modèles.
- Tester rigoureusement la robustesse du modèle face à des situations imprévues ou des tentatives de manipulation.

- Réaliser des simulations d'attaques (*red teaming*) pour valider la sécurité et la fiabilité du modèle.

→ Métriques Clés

- Qualité du contenu : Suivre des métriques comme BLEU, ROUGE, METEOR, pertinence, fluidité, cohérence.
- Qualité des images : Suivre des métriques comme l'Inception Score (IS) et la Frechet Inception Distance (FID).
- Sécurité : Mesurer la présence de contenu toxique ou de biais dans les réponses.
- Fiabilité : Évaluer la fréquence des "hallucinations" et la capacité du modèle à se baser sur des faits (*grounding*).

→ Surveillance Continue (Monitoring)

- Mettre en place une surveillance continue des modèles en production pour détecter la dérive des performances (*model drift*) et les comportements anormaux.
- Configurer des alertes pour les violations de seuils de performance ou de sécurité.
- Utiliser des tableaux de bord pour visualiser les KPIs et suivre l'efficacité du modèle en conditions réelles.

→ Amélioration Continue

- Collecter les données de production et les retours utilisateurs pour ajuster et ré-entraîner les modèles si nécessaire.
- Auditer en temps réel les requêtes et les réponses pour garantir la conformité et la traçabilité.

L'évaluation et la surveillance des modèles ne sont pas des étapes ponctuelles mais des processus continus essentiels à la fiabilité de vos systèmes d'IA. En combinant évaluation

humaine, métriques automatisées et monitoring en temps réel, vous détectez rapidement les dérives de performance et maintenez la qualité de vos résultats. Les données de production et les retours utilisateurs sont vos meilleurs alliés pour améliorer continuellement vos modèles.

Comment mesurer et optimiser la performance de vos modèles ?

Je vous aide à mettre en place un cadre d'évaluation rigoureux et des processus de monitoring adaptés. Une [session de consulting](#) ciblée peut vous permettre de structurer rapidement votre approche d'évaluation.

 Contact : sebastien.chami@blackbygreen.com

 Réserver un échange : <https://calendly.com/sebastien-chami/rdv>

 Formulaire de contact : <https://blackbygreen.fr/contact/>