IMPACT AI

LES BRIEFS DE L'IA RESPONSABLE

7

Cycle de vie et pratiques de l'IA générative

JANVIER 2025

Introduction

Ce « Brief de l'IA responsable » d'Impact AI rend compte de nos travaux et propose des pratiques et des initiatives, afin de garantir que le développement et le déploiement de l'IA générative (GenAI) soient aussi industrialisés et reproductibles que possible dans sa capacité à appliquer des pratiques d'IA responsable à grande échelle.

Il s'appuie sur les retours des membres d'Impact AI et diverses études et initiatives récentes pour livrer un aperçu des stratégies à mettre en œuvre.

1/ Le contexte

L'élan soulevé au cours des vingt-quatre derniers mois par l'IA générative (GenAI) et les modèles de fondation accélère non seulement l'évaluation et l'adoption de l'IA au sein des organisations, mais souligne également la nécessité de mettre en œuvre de nouveaux outils et processus et d'impulser un changement fondamental dans la manière dont les équipes, techniques et non techniques, au sein d'une organisation doivent collaborer pour gérer leurs pratiques de GenAI(responsable) à grande échelle.

Cette nouvelle approche est souvent appelée GenAIOps ou opérations d'IA générative (ou encore LLMOps ou opérations de grands modèles de langage). Il s'agit d'un sous-ensemble de MLOps axé spécifiquement sur l'opérationnalisation et la gestion des grands modèles de langage (LLM) et autres modèles de fondation. Il décrit les pratiques et stratégies opérationnelles de gestion adaptées à tel ou tel modèle jusqu'à sa mise en production et même après son déploiement.
Pour aller un peu plus loin dans le détail, MLOps[1] est un concept qui aligne les personnes, les processus et les plateformes dans l'ensemble de l'organisation, de l'IT aux métiers et à la « Data Science » (DS), afin d'obtenir une valeur métier continue à partir de l'IA du Machine Learning (ML). Il est essentiel que les organisations élaborent leurs propres stratégies en la matière sur la façon dont elles rationalisent les processus et alignent les personnes, afin de s'adapter au mieux à la plateforme MLOps qu'elles retiennent le cas échéant.
Il existe en effet beaucoup de complexités dans la construction de systèmes utilisant des LLM ou d'autres modèles de GenAI. Il faut sélectionner les bons modèles, les affiner aux besoins (fine-tuning) ; concevoir l'architecture et les flux de données associées ; orchestrer les invites (prompts) et l'ancrage (grounding) ; évaluer ces invites pour leur pertinence, les intégrer dans les systèmes et les surveiller en utilisant des chaînes d'outils d'IA responsable, pour ne citer que quelques-unes de ces complexités.

Sans oublier les préoccupations concernant la sûreté, la cybersécurité, la protection des données (sensibles ou à caractère personnel).
Les développeurs manquent d'expérience, d'outils et de processus reproductibles pour évaluer, améliorer et valider les solutions pour délivrer leurs preuves de concept et pour les industrialiser, les mettre à l'échelle et les exploiter en production.
Toutes celles et ceux qui ont déjà commencé leur parcours MLOps verront que les techniques utilisées ouvrent la voie au GenAIOps, comme contrôler les données (d'entraînement et de test), définir les poids, utiliser des chaînes d'outils d'IA responsable pour identifier les biais, les cohortes d'erreurs, et surveiller le système et son ou ses modèles en production. Cependant, contrairement aux modèles de Machine Learning (ML) traditionnels qui ont souvent des résultats plus prévisibles, les modèles GenAI sont par nature non déterministes. Cela oblige à adopter une manière différente de travailler avec eux.
Ainsi, bien que la plupart de ces techniques s'appliquent encore aux systèmes GenAI modernes, il convient d'y ajouter de manière non exhaustive beaucoup d'autres éléments : la configuration de la recherche vectorielle/sémantique, le découpage des données à indexer, l'ingénierie des invites, l'ancrage des données avec des approches comme le RAG (Retrieval-Augmented Generation), les systèmes de sécurité, les évaluations manuelles et automatiques des invites, ou encore la surveillance continue avec des remontées d'alertes automatisées. Toutes ces techniques deviennent les pierres angulaires de bonnes pratiques (cf. ci-après).
À l'instar du MLOps, le GenAIOps est plus qu'une simple adoption de technologies ou de produits. Il s'agit en effet d'organiser la complémentarité entre l'évaluation diligente, la gestion des risques et le déploiement (en temps réel) à grande échelle des systèmes d'IA pour les rendre pleinement opérationnels pour l'entreprise.

1 D'un point de vue technique, le MLOps consiste à appliquer les principes Dev(Sec)Ops aux flux de travail (workflows) ML en utilisant une approche d'intégration continue (CI) aux flux de travail de DS, en automatisant l'entraînement et le processus de test et d'évaluation des modèles et en utilisant la livraison/déploiement continu (CD), pour automatiser le test et le déploiement (en temps réel) des modèles prêts pour la production.

Le retour d'expériences des membres Impact AI

Exxa : réduire le coût financier et environnemental de l'IA générative

Exxa est une startup créée en 2023 et qui s'est fixé pour objectif d'aider les entreprises à développer l'IA générative en en réduisant les coûts financiers et environnementaux. Elle travaille pour des entreprises qui ont beaucoup de données à traiter ou des entreprises spécialisées dans l'IA pour lesquelles Exxa sous-traite ces traitements. Cela consiste tout d'abord à s'assurer, par des expérimentations rapides de modèles d'IA générative et outils génériques, que l'outil est réellement adapté au cas d'usage et s'inscrit de façon pertinente dans un processus métier de l'entreprise. Si cette expérimentation s'avère positive, il peut être souhaitable d'orienter l'entreprise vers des modèles open source plus petits, spécialisés sur des tâches spécifiques ou encore de recourir au traitement différé. L'entreprise n'a pas toujours besoin d'avoir les résultats d'un traitement de données par une IA générative immédiatement. Il est possible de réduire de 10 à 15 fois le coût du traitement et de 100 à 150 fois le coût énergétique, en utilisant les temps de calcul disponibles la nuit, dans des périodes de creux. Et l'on peut aussi favoriser les traitements dans des zones à faible émission en matière énergétique comme la France ou l'Europe du Nord.

Microsoft France : créer une valeur et un impact reproductibles

Microsoft France a mis au point et partagé un cycle de vie structuré pour la gestion, de bout en bout, d'opérations de GenAI avec trois boucles principales (et étapes) à considérer pour : i) rationaliser le déploiement, la gestion et la mise à l'échelle des systèmes GenAI (et de leurs modèles) dans les organisations, et finalement ii) créer une valeur et un impact reproductibles au sein et en dehors d'une organisation à partir de l'investissement en IA :

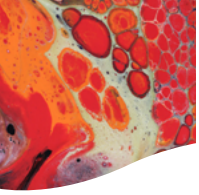
1. Une boucle d'idéation et d'exploration pour identifier un modèle qui s'aligne sur des exigences métier spécifiques. Cela implique d'utiliser un sous-ensemble de données et d'invites (prompts) de base afin d'apprécier les capacités et les limites de chaque modèle à travers :

- → Un prototypage et test de diverses invites, de différentes approches RAG pour la contextualisation, et la validation ou réfutation des hypothèses métier.
- → Des évaluations quant à la pertinence des réponses, d'abord manuelles et pratiques dans une aire de jeu (playground), puis avec des métriques automatisées et des fichiers.

2. Une boucle de construction et d'augmentation pour guider et améliorer le(s) modèle(s) afin de mieux répondre aux besoins spécifiques par :

- → La mise en œuvre de l'approche RAG retenue.
- → L'affinage (fine-tuning) pour aligner les réponses avec les exigences spécifiques de la tâche en ajustant les paramètres du ou des modèles pour le système. (Cette méthode est utilisée lorsque la précision des résultats ne répond pas aux seuils souhaités).
- → Une combinaison des méthodes ci-dessus.

3. Une boucle d'opérationnalisation pour faire passer le système (et le(s) modèle(s) associé(s))



du développement à la production sans qu'il soit nécessaire de mettre en place une infrastructure complexe en :

- → Incorporant des systèmes de sécurité de contenu pour détecter et atténuer les abus et les contenus indésirables, à l'entrée et à la sortie du système, ainsi que dans l'orchestration du ou des modèles et des plug-ins le cas échéant.
- → Intégrant avec des processus d'intégration et de déploiement continus (CI/CD) pour un déploiement du système rationalisé et efficace.
- → Surveillant en continu, pour suivre et optimiser à la fois les performances et la qualité, la sécurité et les risques du système une fois celui-ci en production.

La gestion des risques est cruciale tout au long de ce cycle de vie de bout en bout étant donné la variabilité des résultats des modèles et les risques et préjudices potentiels associés à leur utilisation, et éventuellement du fait de la Loi et des cadres réglementaires applicables, p. ex. pour les systèmes à haut risque selon le Règlement européen sur l'IA, entrée en vigueur le 1er août 2024.

L'infusion d'outils et de pratiques d'IA responsable dans GenAIOps témoigne de la conviction de Microsoft que l'innovation technologique et la gouvernance ne sont pas seulement compatibles, mais se renforcent mutuellement.

Les bonnes pratiques identifiées

Certaines bonnes pratiques comme celles couvertes par l'article What Is LLMOps, Why It Matters & 7 Best Practices In 2025 aident à maintenir les performances, la qualité, la sûreté et la sécurité des systèmes GenAI (et de leurs modèles) tout au long de leur cycle de vie (de développement) :

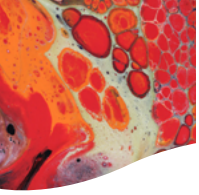
1. Une gestion robuste des données avec des données (d'entraînement) de haute qualité est cruciale pour l'affinage (fine-tuning) des modèles et la mise en œuvre d'approche comme le RAG pour l'ancrage (grounding). La gestion appropriée des données, y compris la collecte, l'exploration, la traçabilité, l'étiquetage et le stockage, etc., est une priorité absolue. L'hygiène des données constitue une partie essentielle de cette pratique dans le cadre d'une nécessaire gouvernance à établir en la matière ; supprimer les enregistrements obsolètes conservés au niveau du stockage réduit p. ex. le risque de fabrications, communément appelées hallucinations, ou faire apparaître des informations incorrectes, est quelque chose à renforcer par une attestation périodique.

2. Une orchestration du développement du système pour planifier et orchestrer soigneusement le processus de développement, y compris l'évaluation dans le contexte et la sélection des modèles, leur affinage, la mise en œuvre du RAG, la définition de messages système, d'invites (prompt) et flux associés, leurs évaluations en matière de performances et de qualité, de risques, de sécurité et de sûreté (safety).

3. Des modalités de déploiement flexible (en temps réel) pour planifier des stratégies qui s'adaptent à différents environnements et cas d'utilisation pour le portefeuille d'actifs GenAI.

4. Une surveillance continue pour suivre les performances des systèmes et modèles GenAI (et d'autres métriques pertinentes) et détecter tout problème ou dérive dans les résultats attendus.

5. Des fonctionnalités de collaboration pour démocratiser l'accès aux modèles GenAI et aux chaînes d'outils et Frameworks d'IA et permettre aux membres d'une équipe pluridisciplinaire étendue (technique et non technique) de travailler ensemble efficacement.



Des projets open source comme Prompt flow représentent des composants pivot essentiels pour assurer une approche structurée, une intégration sans effort avec des chaînes d'outils et des Frameworks d'IA (en open source) comme LangChain, Semantic Kernel, etc.

6. Une gouvernance responsable avec des pratiques bien établies et solides pour garantir une utilisation sûre, sécurisée, digne de confiance et responsable, avec une gestion adéquate des risques, de la conformité aux réglementations comme abordés par les précédents Briefs de l'IA responsable.

7. Une intégration du contexte métier pour combler les lacunes entre les équipes techniques, fonctionnelles et métier afin de garantir que les systèmes et modèles GenAI s'alignent sur les objectifs métier et offrent une valeur continue et reproductible.

Les défis et questions en suspens

Nos membres font les constats suivants :

→ L'évaluation des modèles d'IA générative restent un défi en 2025 avec encore relativement peu de métriques ou de méthodes communes et une instabilité des performances.

→ La gouvernance est indispensable pour gérer la situation évolutive.

→ Une dépendance des utilisateurs aux modèles de fondation utilisés et le besoin de pouvoir comparer et vérifier les modèles avant intégration pour être sûr d'utiliser le bon modèle pour le bon usage.

Pour aller plus loin

• Responsible AI tools and practices in your LLMOps, un article de blog qui couvre l'utilisation des outils et méthodologies d'IA responsable de Microsoft pour aider à atténuer les risques associés aux systèmes GenAI.

• Évaluation du niveau de maturité pour les opérations d'IA générative (GenAIOps), un questionnaire conçu pour aider les organisations à comprendre leurs capacités actuelles et à identifier les domaines à améliorer.

Les résultats de l'évaluation correspondent à un niveau de classement du modèle de maturité GenAIOps, proposant une compréhension générale et un niveau d'application pratique. Ces lignes directrices pragmatiques fournissent aux organisations des liens utiles pour élargir leur base de connaissances GenAIOps.

Remerciements
à Philippe Beraud (Microsoft), Emilie Sirvent Hien (Orange)
et Matthieu Capron (Crédit Agricole) pour leur engagement
et leur contribution essentielle à ces Briefs de l'IA responsable.
à tous les participants aux ateliers qui ont bien voulu
partager leurs expériences et cas d'usage.

Retrouvez-nous sur

www.impact-ai.fr

pour rejoindre nos initiatives et retrouver nos divers travaux !


IMPACT AI