Guardrails pour les LLMs : garantir une génération alignée et sûre

Données traitées	Outils
Texte	Tout LLM
Public cible	**Mise en œuvre**
Initiateur de projet	€€€
Gain	**Effort**
Confiance dans l'IA	
Pré requis	Avoir un projet d'IA générative sur mesure à lancer

Contributeur
Jean-Baptiste Juin
Co-fondateur et directeur R&D
chez Cross Data

Cross Data conçoit, construit, déploie
et maintien des algorithmes et des IAs
pour améliorer la performance de ses clients.

PROBLÉMATIQUE

La technologie des LLMs (IA générative pour le langage) est intrinsèquement probabiliste. Les textes générés peuvent donc parfois présenter du contenu inapproprié : contenu offensant, fausses affirmations, etc.

Dans le contexte d'un outil d'IA générative pour l'entreprise ce comportement peut se traduire par des textes générés : qui dé-servent l'image de l'entreprise, qui exposent des données internes, qui ne sont pas alignés avec les valeurs de l'organisation, etc.

Il est donc essentiel de prendre en compte, lors de l'expression de son besoin dans un tel outil, d'exprimer clairement l'alignement attendu ou de soulever le problème avec le concepteur de la solution pour construire, durant des ateliers préparatoires, cette « ligne éditoriale ».

ANALYSER ET EXPRIMER LES BESOINS

La première étape essentielle est de bien comprendre les besoins de son organisation en terme de « ligne éditoriale » pour l'IA générative.

- Quels sont les « interdits », les sujets à ne surtout pas aborder ?
- Quel degré minimal de qualité est attendu pour la génération finale ?

Source : Image générée par mistral.ai

Guardrails pour les LLMs : garantir une génération alignée et sûre

On constate généralement qu'une génération automatique doit être évaluée selon plusieurs « axes », parfois génériques (complétude de la réponse, exactitude des affirmations et respect du contexte principalement) mais souvent spécifique au métier sous-jacent ou à la tâche demandée (diversité des réponses, respect du style linguistique d'une marque, etc). La mesure de la qualité n'est donc souvent pas « binaire » mais est à adapter en fonction du projet en choisissant quel niveau on cherche à atteindre sur les différents axes définis pour le projet.

Le travail de définition de ce besoin est, en général, à faire avec le prestataire de la solution qui pourra faire émerger, en fonction de la situation : les différents axes d'évaluation, les « interdits », etc. Tous ces éléments seront ensuite à traduire en « guardrails » pour éviter les « sorties de route » de l'IA générative.

MISE EN ŒUVRE

La mise en place de ces « guardrails » passe par une diversité de méthodes, chacune ayant des avantages et inconvénients en fonction du cas à traiter. On distingue 2 grandes catégories de techniques pour construire des garde-fous : les vérifications syntaxiques et les vérifications sémantiques.

Les différentes méthodes techniques utilisées vont avoir un impact plus ou moins grand sur le temps de latence et/ou sur le temps de génération total. L'usage de ces techniques nécessite donc de faire un compromis entre expérience utilisateur et besoin de contrôle.

Il existe des outils ou des librairies à la fois propriétaire ou open-source pour implémenter des « guardrails ». La mise en place requiert souvent une étape d'expérimentation car chacune de ces solutions propose des fonctionnalités et des performances qui peuvent varier.

POUR ALLER PLUS LOIN

llamaguard (LLM dédié par Meta)

NemoGuardrails (librairie par Nvidia)

deepeval (librairie proposant des guardails sous forme de tests)