Les principaux risques :

Divulgation de données sensibles
Prompt Injection
Exécution de commandes malveillantes
Traitement non sécurisé des sorties

Les mitigations :

Filtrage des entrées
Validation de schéma
Durcissement du system policy
Limites d'actions et revues humaines

SECURITÉ DES LLMS

GABIN MERLOT-DIMET

LUCAS GOUBET