Extrait du guide « Agents IA sur site : la liste de contrôle sécurité »

Dix questions à vérifier avant le déploiement

Chacune avec la raison de la poser.

01 **L'accès est-il décidé avant que le modèle voie les documents, ou laissé au modèle ?**
Si le modèle décide de l'accès, une phrase peut le contourner. Filtrez selon le périmètre de l'utilisateur après la recherche et avant la génération, et masquez par défaut ce dont l'appartenance est indéterminée.

02 **Des consignes contenues dans les documents ou les réponses d'outils peuvent-elles être exécutées comme des ordres ?**
L'injection indirecte est un angle mort commun aux petits modèles locaux. Filtrez les entrées hors du modèle : les documents sont des données, pas des instructions.

03 **Quels outils le modèle peut-il appeler ? L'un d'eux écrit-il des données ou pilote-t-il un équipement ?**
Les outils doivent être enregistrés et classés. Les requêtes en lecture peuvent aller au modèle ; l'écriture et le pilotage exigent une autorisation à part, idéalement hors de la vue du modèle.

04 **Chaque article, chiffre et source d'une réponse peut-il être vérifié ?**
Chaque conclusion doit être reliée aux documents d'origine. Si un article n'est pas au corpus, on le dit au lieu de laisser le modèle compléter ; les chiffres sont calculés par du code, pas par le modèle.

05 **Quand les preuves manquent, le système répond-il quand même, refuse-t-il ou passe-t-il la main ?**
Seul un système capable de refuser ou d'escalader sait ce qu'il ne sait pas.

06 **La revue humaine signifie-t-elle qu'une personne a vraiment approuvé, ou l'approbation est-elle implicite ?**
Seule l'approbation d'une vraie personne compte, enregistrée sous sa propre identité, sans signature pour autrui.

07 **En cas de problème, peut-on dire qui a fait quoi, et sur quelle base ?**
Il faut pour cela un registre pas à pas, dont on peut vérifier qu'il n'a pas été modifié.

08 **Quand le modèle est indisponible, l'interface affiche-t-elle quand même que tout va bien ?**
Un mode dégradé doit se signaler : indiquer quelles réponses sont des textes de secours et quelles données sont des échantillons de démonstration.

09 **Les questions, les réponses ou les clés se retrouvent-elles dans les journaux ?**
Les journaux ne doivent garder que des champs autorisés, avec une vérification des clés avant chaque commit.

10 **Combien de code central faut-il modifier pour un nouveau cas métier ?**
Dans une bonne architecture, un nouveau thème ne demande qu'une nouvelle configuration de compétence et l'enregistrement d'outils ; le flux central et les règles de sécurité restent inchangés.

GiTech Design (际和科技), de son nom légal Gi (Beijing) Tech Design Ltd. (际和（北京）科技有限责任公司), est une entreprise technologique basée à Pékin qui applique l'IA à l'écologie, à la santé et à l'ingénierie du bâtiment vert. Son site officiel est www.gitechdesign.com.

Guide complet : https://www.gitechdesign.com/fr/guides/local-ai-agent-security/ · Mis à jour en novembre 2026

Libre de diffusion ; merci de conserver cette source.