RISQUES CYBER
- FOCUS IA

Mémo sur les sujets cyber de l'application Isiadoc du point de vue IA

Version	Modification	Fait par	Date
1.0	Initialisation	Pier-Jean Malandrino	21/11/2025

1. De quoi parle t-on ? **3**
1.1. Explication 3
1.2. Détail technique 3
2. Nos mises en oeuvre sécurités technique **5**
2.1. Isolation du LLM sur un serveur spécifique 5
2.1.1. Conformité 5
2.1.2. Sécurité 5
2.2. Isolation fonctionnelle du LLM 5
2.2.1. Conformité 5
2.2.2. Sécurité 5
2.3. Nettoyage des outputs 6
2.3.1. Conformité 6
2.3.2. Sécurité 6
2.4. Contrôle des inputs via templating 6
2.4.1. Conformité 6
2.4.2. Sécurité 6
2.5. Authentification et autorisation RBAC via OAuth2 7
2.5.1. Conformité 7
2.5.2. Sécurité 7
3. Les garanties pour nos clients **8**
4. Sources **9**

1. De quoi parle t-on ?

1.1. Explication

Dans le contexte spécifique d'Isiadoc, les risques de cybersécurité liés à l'intelligence artificielle sont fondamentalement **différents de ceux évoqués dans les médias ou les cas d'usage grand public**. Notre système est conçu pour répondre à des questions techniques sur des documentations de piscines et équipements aquatiques. **Il s'agit d'un assistant documentaire, pas d'un agent autonome capable d'actions.**

Le risque principal souvent évoqué, **le "prompt injection", consiste à tenter de faire dévier l'IA de son comportement prévu en lui injectant des instructions malveillantes.** Dans les systèmes où l'IA peut exécuter des commandes, accéder à des bases de données sensibles ou déclencher des actions, ce risque est critique. Un attaquant pourrait théoriquement faire exfiltrer des données, exécuter du code malveillant ou compromettre l'intégrité du système.

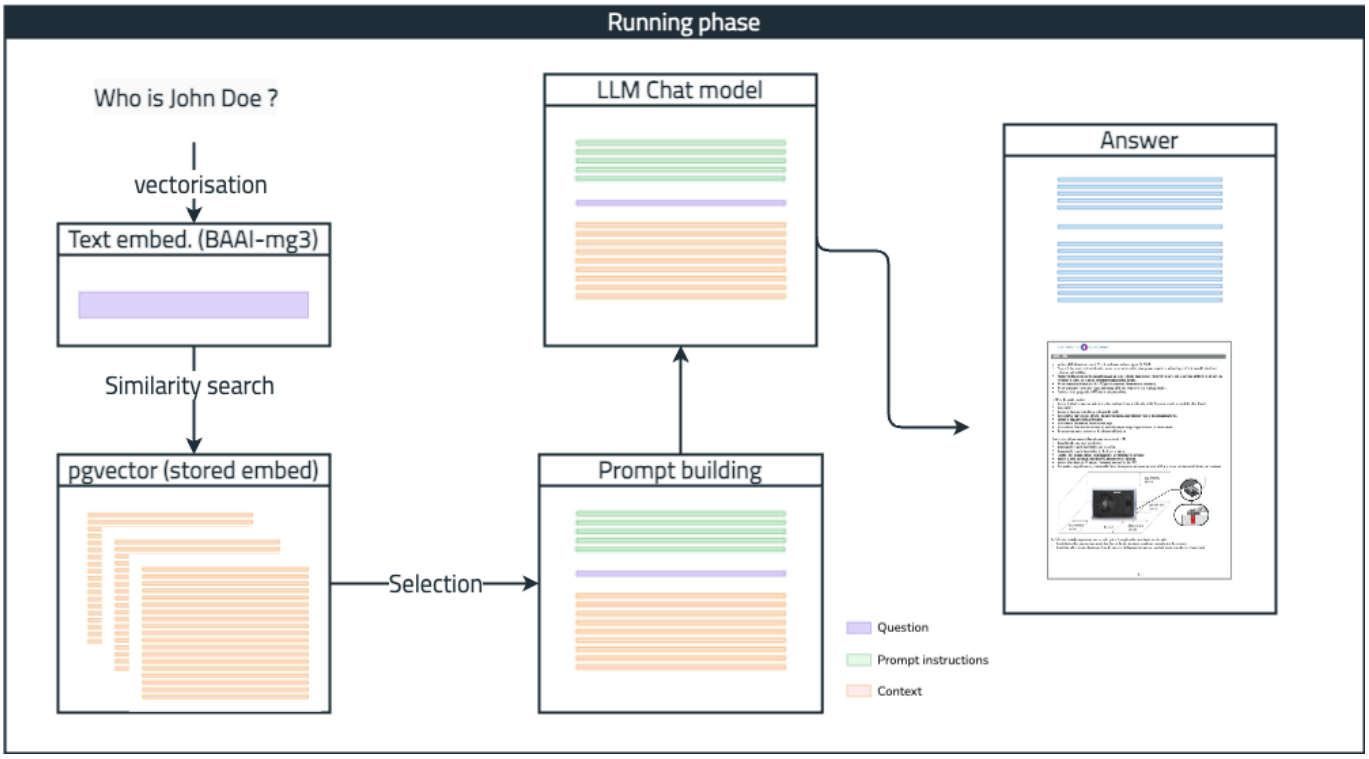
Cependant, dans notre architecture Isiadoc, **ce risque est structurellement neutralisé**. Notre IA fonctionne comme un moteur de recherche avancé qui lit des documents techniques et formule des réponses textuelles. Elle n'a aucune capacité d'action au-delà de la génération de texte. **Si un utilisateur tentait une injection malveillante, le pire scénario serait une réponse incohérente ou incorrecte affichée à l'écran, sans aucune conséquence sur les systèmes ou les données.**

Les véritables enjeux de sécurité pour Isiadoc se situent ailleurs. La protection de la propriété intellectuelle contenue dans les documentations techniques constitue notre priorité. Il s'agit **d'empêcher l'extraction massive de notre base documentaire** par des concurrents ou des acteurs malveillants. **La qualité et la fiabilité des réponses** représentent également un enjeu crucial, car une mauvaise information technique pourrait avoir des conséquences sur les installations de nos clients. Enfin, **la conformité réglementaire, notamment le respect du RGPD/AI Act dans le traitement des questions des utilisateurs**, reste une obligation légale incontournable.

1.2. Détail technique

L'architecture Isiadoc repose sur une approche RAG (Retrieval-Augmented Generation) qui limite intrinsèquement la surface d'attaque. Le flux de données suit un pipeline unidirectionnel strict :

1. **Réception de la requête** utilisateur via l'interface Flutter
2. **Authentification et autorisation** via Auth0 (OAuth2)
3. **Vectorisation** de la requête via le modèle d'embed
4. **Recherche de similarité** dans pgvector (PostgreSQL)
5. **Enrichissement du contexte** avec les chunks documentaires pertinents
6. **Génération de la réponse** par le modèle LLM de chat
7. **Sanitisation** et transmission au frontend



Ce pipeline ne permet aucun "callback" vers les systèmes sources. Le modèle LLM fonctionne en mode stateless, sans mémoire persistante entre les requêtes, et sans aucune capacité d'appel système ou d'accès réseau. Cette architecture répond directement aux vulnérabilités LLM01 (Prompt Injection), LLM06 (Excessive Agency) et LLM05 (Improper Output Handling) du référentiel OWASP Top 10 for LLM Applications 2025.

2. Nos mises en oeuvre sécurités technique

2.1. Isolation du LLM sur un serveur spécifique

Notre modèle d'intelligence artificielle est déployé sur une infrastructure dédiée, strictement séparée des systèmes de production et des bases de données clients. Cette isolation physique garantit qu'aucune compromission du LLM ne peut affecter les autres composants du système, assurant ainsi une barrière de sécurité fondamentale. La segmentation réseau et la conteneurisation Docker/Ollama assurent le confinement des composants, limitant les interactions aux seules communications strictement nécessaires.

2.1.1. Conformité

Conformément à l'article 15 du AI Act qui exige des garanties de robustesse et de sécurité afin d'éviter tout impact non maîtrisé sur d'autres systèmes, cette architecture isolée garantit un confinement strict et réduit considérablement le risque de compromission croisée [1].

2.1.2. Sécurité

Sur le plan de la sécurité applicative, cette isolation répond principalement au risque LLM01 identifié dans le référentiel OWASP Top 10 2025, qui concerne les attaques par prompt injection [2]. En cloisonnant l'environnement de déploiement, nous diminuons l'impact d'une éventuelle manipulation malicieuse des entrées du LLM et limitons la propagation hors de son périmètre. Cette mesure complète aussi la gestion du risque LLM03, relatif à la vulnérabilité de la chaîne d'approvisionnement, par la limitation stricte des surfaces d'attaque externes [2].

2.2. Isolation fonctionnelle du LLM

Le modèle d'IA opère strictement en mode "question-réponse". Il reçoit des questions enrichies à l'aide de notre système de récupération augmentée de documents (RAG), ce qui lui permet de contextualiser et de générer des réponses adaptées. Il ne possède aucune capacité d'exécuter des appels API, d'accéder aux systèmes de fichiers ou d'interagir avec d'autres services. Cette limitation fonctionnelle radicale restreint fortement l'autonomie du modèle, évitant ainsi les risques liés à des décisions ou actions non supervisées pouvant impacter le système ou l'environnement. En pratique, seul le texte généré basé sur les données fournies est retourné, sans aucune fonction annexée ou activation d'opérations externes.

2.2.1. Conformité

L'article 14 du Règlement (UE) 2024/1689 relatif à l'AI Act impose que les systèmes d'intelligence artificielle incorporent une gouvernance humaine solide, notamment en limitant les prises de décision automatiques non contrôlées [1]. Le mode strict "question-réponse" garantit cette supervision humaine implicite, empêchant l'IA d'exécuter de manière autonome des actions à risque sans intervention ou validation humaine. Cette conformité réduit les risques réglementaires liés à un débordement fonctionnel non maîtrisé.

2.2.2. Sécurité

Du point de vue de la sécurité applicative, cette architecture répond spécifiquement aux risques identifiés dans le référentiel OWASP Top 10 pour les LLM 2025. Elle élimine le risque LLM06 dit d'« excessive agency », où un LLM pourrait bénéficier de permissions excessives et entreprendre des actions potentiellement dommageables sans supervision [2]. De même, elle prévient le risque LLM07 de fuite des prompts système, grâce à l'absence d'interactions qui pourraient exposer des instructions ou configurations sensibles. Ces mesures s'appuient sur

la limitation stricte des capacités du LLM et une architecture pensée pour une surface d'attaque minimale par un modèle cloisonné et contrôlé [2].

2.3. Nettoyage des outputs

Les réponses générées par le LLM sont systématiquement traitées comme du texte brut et non-exécutable afin d'éviter tout risque d'exécution de code malveillant côté client. Pour cela, les sorties subissent un processus complet d'échappement HTML qui neutralise toute balise ou instruction potentiellement dangereuse. Ce traitement est équivalent à afficher un document texte classique, même lorsque la réponse contient des éléments semblant être des instructions ou du code. Ainsi, aucune interprétation ni exécution ne peut survenir, garantissant que le contenu affiché reste strictement informatif et sécurisé.

2.3.1. Conformité

Conformément à l'article 15 du AI Act, les systèmes d'IA doivent impérativement garantir que leurs sorties ne soient pas source d'altérations malveillantes ni de comportements dangereux non anticipés [1]. Ce mécanisme de sanitisation constitue donc une mesure essentielle pour satisfaire aux exigences de robustesse et de cybersécurité formulées dans ce cadre réglementaire.

2.3.2. Sécurité

Du point de vue de la cybersécurité, cette pratique répond précisément au risque identifié par OWASP sous la référence LLM05, qui porte sur la mauvaise gestion des sorties de LLM [2]. Ce risque inclut entre autres des vulnérabilités telles que l'injection de scripts (XSS) ou des données malveillantes pouvant compromettre le système ou les utilisateurs. Notre pipeline d'assainissement permet ainsi de garantir la solidité et l'intégrité des données retournées, limitant efficacement toute possibilité d'exploitation via les sorties du modèle.

2.4. Contrôle des inputs via templating

Chaque question utilisateur intégrée dans notre système est encadrée à travers un template de prompt prédéfini, rigide et non modifiable. Ce template délimite explicitement le périmètre des réponses attendues du LLM, en incluant les instructions système qui orientent le comportement du modèle, ainsi que le ton et les règles à respecter, sans possibilité pour l'utilisateur de modifier quoi que ce soit sauf la zone spécifique de son questionnement. Cette structuration garantit une normalisation stricte des entrées, limitant drastiquement les risques liés à des manipulations malveillantes ou involontaires des prompts. Le système impose également des contrôles complémentaires, comme la validation syntactique et des filtres sur les patterns d'entrée, protégeant ainsi la qualité et la sécurité des données soumises au modèle.

2.4.1. Conformité

Cette approche s'inscrit parfaitement dans les exigences de l'article 10 du Règlement (UE) 2024/1689 (AI Act) qui préconise une gestion structurée, rigoureuse et vérifiée des données d'entrée aux systèmes d'IA, assurant la qualité, la pertinence et la sécurité des inputs [1]. En encadrant strictement les questions via un templating, nous respectons les bonnes pratiques imposées pour limiter la surface d'attaque, garantir la responsabilité et prévenir les entrées non conformes ou non sécurisées.

2.4.2. Sécurité

Sur le plan de la sécurité, cet encadrement constitue une des méthodes recommandées par le référentiel OWASP Top 10 pour les LLM 2025 pour mitiger le risque dit LLM01 – Prompt Injection, qui est la vulnérabilité la plus courante et critique [2]. Le verrouillage total des instructions système et la limitation à la simple zone question empêchent que des entrées malicieuses ne soient intégrées dans le prompt système, réduisant ainsi les surfaces d'intrusion par injection de commande ou manipulation du comportement du LLM. De plus, les

contrôles de validation et filtres renforcent la défense en profondeur face aux risques d'attaques directes ou indirectes visant à altérer les réponses du modèle.

2.5. Authentification et autorisation RBAC via OAuth2

L'accès au système est sécurisé par une solution d'authentification standard OAuth2, mise en œuvre via la plateforme Auth0 [3]. Chaque utilisateur doit obligatoirement s'authentifier par ce biais, garantissant ainsi son identité de manière fiable et sécurisée. L'architecture RBAC (Role-Based Access Control) intégrée permet de définir des rôles précis pour chaque utilisateur, ce qui autorise un contrôle granulaire des droits d'accès aux différentes ressources documentaires et fonctionnalités [4]. Ce modèle est finement paramétré dans Auth0, via des règles et politiques qui peuvent assigner dynamiquement des permissions selon le contexte métier. Par ailleurs, des quotas d'utilisation sont appliqués pour éviter toute surcharge ou abus, et l'ensemble des accès est tracé et historisé strictement, permettant une auditabilité complète et un suivi en temps réel via des journaux sécurisés.

2.5.1. Conformité

Dans le cadre de l'article 14 du AI Act, la gouvernance humaine implique que les systèmes IA disposent d'un contrôle rigoureux et continu des accès afin de prévenir tout usage abusif ou non autorisé [1]. Notre architecture répond pleinement à cette exigence par l'intégration d'une solution mature et éprouvée d'authentification et d'autorisation, assurant une gestion claire, traçable et réversible des droits utilisateurs. Ce dispositif incarne une gouvernance humaine effective, assurant la responsabilisation et la protection des données dans des environnements à risque limité.

2.5.2. Sécurité

D'un point de vue cybersécurité, l'usage d'Auth0 allié au modèle RBAC limite grandement les risques LLM02 (divulgation d'informations sensibles) et LLM10 (consommation excessive) répertoriés par le cadre OWASP Top 10 pour LLMs [2]. En restreignant strictement l'accès en fonction des rôles et des permissions, les fuites d'informations critiques sont évitées. De plus, la gestion des quotas assure que les ressources ne peuvent être accaparées par des usages excessifs ou malveillants, prévenant ainsi les menaces de déni de service. Les logs d'accès sécurisés et immuables permettent une traçabilité fine et la détection rapide d'incidents.

3. Les garanties pour nos clients

Notre approche globale se traduit par trois garanties majeures qui définissent notre différenciation sur le marché. Premièrement, une sécurité by-design pensée dès la conception : Isiadoc limite volontairement ses fonctionnalités à la fourniture d'informations techniques sûres, excluant d'emblée toute action non maîtrisée ou modification des données clients, ce qui limite les surfaces de risques.

Deuxièmement, l'auditabilité et la traçabilité sont au cœur du système, avec un logging exhaustif de chaque interaction, des contrôles d'accès rigoureux et un tableau de bord dédié pour le suivi en temps réel. Cette transparence répond aux tendances actuelles de gouvernance et garantit un contrôle plein et entier aux clients.

Enfin, la conformité intégrée RGPD et AI Act structure nos engagements juridiques et techniques. La minimisation des données, la gestion des droits d'accès, l'effacement, ainsi que la robustesse technique sécurisent la conformité réglementaire, garantissant à nos clients une tranquillité juridique ajoutée à la sécurité opérationnelle.

En synthèse, choisir Isiadoc c'est choisir la sécurité par la simplicité. Pendant que d'autres multiplient les fonctionnalités et les risques associés, nous perfectionnons une seule chose : vous donner accès à votre documentation technique de manière sûre, fiable et conforme. C'est cette focalisation qui fait notre force et garantit votre tranquillité d'esprit.

4. Sources

[1] : Règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 relatif à l'intelligence artificielle (AI Act). Journal officiel de l'Union européenne, L 168/1, 13 juin 2024. Consulté le 21 novembre 2025, sur https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX%3A32024R1689

[2] : OWASP Foundation. (2025). OWASP Top 10 for Large Language Model (LLM) Applications 2025. Consulté le 21 novembre 2025, sur https://owasp.org/www-project-top-10-for-large-language-model-applications/

[3] : Hardt, D. (2012). RFC 6749 - The OAuth 2.0 Authorization Framework. Internet Engineering Task Force. Consulté le 21 novembre 2025, sur https://datatracker.ietf.org/doc/html/rfc6749

[4] : National Institute of Standards and Technology (NIST). (2020). NIST Special Publication 800-53 Revision 5: Security and Privacy Controls for Information Systems and Organizations. Consulté le 21 novembre 2025, sur https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r5.pdf