Comparaison architecturale : IA standard / RAG et COMPAiSS

Architecture IA standard à génération d'abord / RAG			Architecture COMPAiSS à inférence verrouillée par autorisation		
ÉTAPE	**CE QUI SE PRODUIT À CETTE ÉTAPE**	**CE QUE CELA IMPLIQUE POUR L'EXACTITUDE ET LE RISQUE**	**ÉTAPE**	**CE QUI SE PRODUIT À CETTE ÉTAPE**	**CE QUE CELA IMPLIQUE POUR L'EXACTITUDE ET LE RISQUE**
Inférence du modèle (toujours ouverte)	Le modèle d'IA est actif et prêt à générer des réponses par défaut. L'inférence commence dès qu'une question est reçue.	*Le système présume qu'il fournira une réponse avant même de savoir si cela est approprié.*	**Inférence du modèle (toujours ouverte)**	Le modèle d'IA n'est pas actif par défaut. L'inférence est verrouillée et ne s'exécutera que si l'autorisation est accordée.	*Le système ne présume pas qu'il fournira une réponse avant d'avoir déterminé si cela est approprié.*
Réception de la question	La question de l'utilisateur est reçue dans n'importe quelle langue et transmise au système pour traitement.	*Aucune vérification du périmètre institutionnel ni de l'autorisation n'est effectuée à cette étape.*	**Réception de la question**	La question de l'utilisateur est reçue dans n'importe quelle langue et immédiatement évaluée quant au périmètre institutionnel et à l'autorisation.	*Les questions hors du périmètre institutionnel autorisé sont bloquées avant tout raisonnement de l'IA.*
Traduction et normalisation linguistique	Si la question n'est pas en anglais, elle peut être traduite dans le cadre du processus de récupération et de raisonnement.	*La traduction peut influencer les documents récupérés et la façon dont le sens est interprété.*	**Traduction et normalisation linguistique**	Si la question n'est pas en anglais, elle est traduite uniquement pour normaliser le sens. Il s'agit d'une étape strictement linguistique, sans récupération ni raisonnement.	*La traduction n'influence pas le choix des sources ni la façon dont les réponses sont déterminées.*
Génération d'incorporation	La question est convertie en représentation mathématique afin de permettre une recherche par similarité parmi les documents stockés.	*Le système effectue une recherche générale dans l'ensemble des données disponibles, sans filtrage institutionnel.*	**Génération d'incorporation**	La question normalisée est convertie en représentation mathématique, mais uniquement à l'intérieur d'un espace de vérité institutionnel prédéfini et délimité.	*La recherche est restreinte aux sources institutionnelles de la liste verte approuvées avant tout raisonnement de l'IA.*
Récupération vectorielle / documentaire (RAG)	Le système récupère les documents ou passages jugés pertinents selon la correspondance par similarité. Ceux-ci peuvent être externes, périmés ou non autorisés.	*Le contenu récupéré peut ne pas être vérifié ni autorisé par l'institution.*	**Récupération vectorielle / documentaire (RAG)**	Le système récupère uniquement à partir de sources institutionnelles autorisées. Si aucun document faisant autorité n'existe, le processus s'arrête ici.	*Seul un contenu vérifié et approuvé par l'institution peut servir de base à une réponse.*
Assemblage du prompt	L'IA combine la question de l'utilisateur et les documents récupérés pour construire un prompt destiné à la génération de la réponse.	*L'IA peut combler des lacunes, inférer du contexte ou s'appuyer sur des données d'entraînement au-delà de ce qui a été récupéré.*	**Assemblage du prompt**	Si l'autorisation est accordée, l'IA est instanciée et combine la question avec les documents autorisés. L'IA ne peut inférer au-delà de ce qui est explicitement fourni.	*L'IA est limitée aux sources institutionnelles et ne peut combler des lacunes à partir de données d'entraînement générales.*
Contrôles post-génération	Une fois la réponse générée, elle peut être soumise à des filtres, des vérifications de sécurité, des couches de modération ou une révision humaine afin de corriger les erreurs.	*Ces contrôles réduisent les risques, mais n'éliminent pas les hallucinations ni les affirmations non étayées.*	**Contrôles post-génération**	Étant donné que la réponse repose exclusivement sur des sources autorisées, le filtrage post-génération est superflu pour assurer l'exactitude et l'autorité.	*Les hallucinations fondées sur des sources non autorisées sont structurellement empêchées et non simplement détectées après coup.*
Réponse finale transmise	La réponse est retraduite (au besoin) et présentée à l'utilisateur, souvent accompagnée de mentions sur les inexactitudes possibles.	*Les utilisateurs peuvent recevoir des réponses contenant des liens non autorisés, des informations non vérifiées ou des sources mixtes.*	**Réponse finale transmise**	La réponse est retraduite (au besoin) et présentée à l'utilisateur, fondée exclusivement sur les documents institutionnels. Si aucune réponse n'est autorisée, l'utilisateur en est informé.	*Les utilisateurs reçoivent uniquement des réponses précises et autorisées, ou un message clair indiquant qu'aucune réponse n'est disponible.*
Dans ce modèle, l'IA est ouverte dès le départ et les réponses sont contrôlées par étapes, mais les hallucinations ne sont pas éliminées.			Dans ce modèle, le raisonnement de l'IA est verrouillé avant même de commencer, ce qui empêche la génération de toute réponse non étayée.		