

# Comparaison architecturale : IA/RAG standard vs COMPAiSS

## Architecture IA/RAG Standard Axée sur la Génération

ÉTAPE	CE QUI SE PRODUIT À CETTE ÉTAPE	CE QUE CELA SIGNIFIE POUR L'EXACTITUDE ET LE RISQUE
<b>Inférence du modèle (toujours ouverte)</b>	Le modèle d'IA est actif et prêt à générer des réponses par défaut. L'inférence commence dès qu'une question est reçue.	<i>Le système présume qu'il fournira une réponse avant de savoir si une réponse est appropriée.</i>
<b>Réception de la question de l'utilisateur</b>	La question de l'utilisateur est reçue dans n'importe quelle langue et transmise au système pour traitement.	<i>Aucune vérification de la portée institutionnelle ou de l'autorisation n'a lieu à ce stade.</i>
<b>Traduction / normalisation linguistique</b>	Si la question n'est pas en anglais, elle peut être traduite dans le cadre du processus de récupération et de raisonnement.	<i>La traduction peut influencer quels documents sont récupérés et comment le sens est interprété.</i>
<b>Génération de plongements</b>	La question est convertie en une représentation mathématique pour permettre la recherche de similitude dans les documents stockés.	<i>Le système effectue une recherche large dans toutes les données disponibles sans filtrage institutionnel.</i>
<b>Récupération vectorielle / documentaire (RAG)</b>	Le système récupère des documents ou des passages qui semblent pertinents selon la correspondance de similitude. Cela peut inclure des sources externes, désuètes ou non autorisées.	<i>Le contenu récupéré peut ne pas être vérifié ou autorisé par l'institution.</i>
<b>Assemblage de l'invite</b>	L'IA combine la question de l'utilisateur avec les documents récupérés pour construire une invite de génération de réponse.	<i>L'IA peut combler des lacunes, inférer le contexte ou puiser dans les données</i>

ÉTAPE	CE QUI SE PRODUIT À CETTE ÉTAPE	CE QUE CELA SIGNIFIE POUR L'EXACTITUDE ET LE RISQUE
		<i>d'entraînement au-delà de ce qui a été récupéré.</i>
Contrôles post-génération	Après la génération de la réponse, celle-ci peut passer par des filtres, des vérifications de sécurité, des couches de modération ou un examen humain pour détecter les erreurs.	<i>Les contrôles réduisent mais n'éliminent pas les hallucinations ou les affirmations non fondées.</i>
Livraison de la réponse finale	La réponse est traduite (si nécessaire) et présentée à l'utilisateur, souvent avec des avertissements concernant l'inexactitude potentielle.	<i>Les utilisateurs reçoivent des réponses qui peuvent inclure des liens non autorisés, des informations non vérifiées ou des sources mixtes.</i>
Dans ce modèle, l'IA est ouverte dès le départ et les réponses sont contrôlées par étapes, mais les hallucinations ne sont pas éliminées.		

## Architecture COMPAiSS à Exécution Contrôlée

ÉTAPE	CE QUI SE PRODUIT À CETTE ÉTAPE	CE QUE CELA SIGNIFIE POUR L'EXACTITUDE ET LE RISQUE
<b>Inférence du modèle (toujours ouverte)</b>	Le modèle d'IA n'est pas actif par défaut. L'inférence est contrôlée et ne se produira que si l'autorisation est accordée.	<i>Le système ne présume pas qu'il répondra avant de déterminer s'il devrait le faire.</i>
<b>Réception de la question de l'utilisateur</b>	La question de l'utilisateur est reçue dans n'importe quelle langue et immédiatement évaluée pour la portée institutionnelle et l'autorisation.	<i>Les questions hors de l'autorité institutionnelle sont bloquées avant le début de tout raisonnement par l'IA.</i>
<b>Traduction / normalisation linguistique</b>	Si la question n'est pas en anglais, elle est traduite uniquement pour normaliser le sens. Il s'agit d'une étape linguistique seulement, non une étape de récupération ou de raisonnement.	<i>La traduction n'influence pas quelles sources sont considérées ou comment les réponses sont déterminées.</i>
<b>Génération de plongements</b>	La question normalisée est convertie en une représentation mathématique, mais seulement à l'intérieur d'un espace de vérité institutionnel prédéfini.	<i>La recherche est limitée aux sources approuvées par l'institution et inscrites sur la liste verte avant tout raisonnement par l'IA.</i>
<b>Récupération vectorielle / documentaire (RAG)</b>	Le système récupère uniquement à partir de sources institutionnelles autorisées. Si aucun document faisant autorité n'existe, le processus s'arrête ici.	<i>Seul le contenu vérifié et approuvé par l'institution peut informer une réponse.</i>
<b>Assemblage de l'invite</b>	Si l'autorisation est accordée, l'IA est instanciée et combine la question avec les documents autorisés. L'IA ne peut inférer au-delà	<i>L'IA est confinée aux sources institutionnelles et ne peut pas combler</i>

ÉTAPE	CE QUI SE PRODUIT À CETTE ÉTAPE	CE QUE CELA SIGNIFIE POUR L'EXACTITUDE ET LE RISQUE
	de ce qui est explicitement fourni.	<i>les lacunes à partir des données d'entraînement générales.</i>
<b>Contrôles post-génération</b>	Puisque la réponse est basée uniquement sur des sources autorisées, le filtrage post-génération n'est pas nécessaire pour l'exactitude ou l'autorité.	<i>Les hallucinations basées sur des sources non autorisées sont structurellement prévenues, non détectées rétroactivement.</i>
<b>Livraison de la réponse finale</b>	La réponse est traduite (si nécessaire) et présentée à l'utilisateur, fondée exclusivement sur les documents institutionnels. Si aucune réponse n'a été autorisée, l'utilisateur en est informé.	<i>Les utilisateurs reçoivent uniquement des réponses précises et autorisées ou un énoncé clair qu'aucune réponse n'est disponible.</i>
<b>Dans ce modèle, le raisonnement par l'IA est contrôlé avant qu'il ne commence, empêchant la génération de réponses non fondées.</b>		