Comparaison architecturale : IA/RAG standard vs COMPAiSS

Architecture IA génération-d'abord standard / RAG

ÉTAPE	CE QUI SE PRODUIT À CETTE ÉTAPE	CE QUE CELA IMPLIQUE POUR LA PRÉCISION 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 suppose qu'il fournira une réponse avant de savoir si celle-ci est appropriée.
Réception de la question de l'utilisateur	La question de l'utilisateur est reçue dans n'importe quelle langue et transmise au système pour traitement.	Aucun contrôle de portée institutionnelle ni d'autorisation n'a lieu à cette étape.
Traduction / Normalisation linguistique	Si la question n'est pas en anglais, elle peut être traduite dans le cadre du processus de recherche et de raisonnement.	La traduction peut influencer les documents récupérés et l'interprétation du sens.
Génération d'embeddings	La question est convertie en une représentation mathématique afin de permettre une recherche par similarité dans les documents stockés.	Le système effectue une recherche large à travers toutes les données disponibles sans filtrage institutionnel.
Récupération vectorielle / de documents (RAG)	Le système récupère des documents ou passages jugés pertinents selon une correspondance par similarité. Cela peut inclure des sources externes, désuètes ou non autorisées.	Le contenu récupéré peut ne pas être vérifié ni autorisé au niveau institutionnel.
Assemblage du prompt	L'IA combine la question de l'utilisateur avec les documents récupérés afin de construire un prompt pour générer une réponse.	L'IA peut combler des lacunes, inférer du contexte ou puiser dans ses données d'entraînement au-delà de ce qui a été récupéré.
Contrôles post-génération	Une fois la réponse générée, elle peut passer par des filtres, vérifications de sécurité, couches de modération ou une révision humaine afin de détecter des erreurs.	Les contrôles réduisent, mais n'éliminent pas, les hallucinations ou affirmations non étayées.
Réponse finale livrée	La réponse est retraduite (au besoin) et présentée à l'utilisateur, souvent accompagnée d'avertissements concernant une possible inexactitude.	Les utilisateurs reçoivent des réponses pouvant inclure des liens non autorisés, des informations non vérifiées ou des sources mixtes.

ÉTAPE	CE QUI SE PRODUIT À CETTE ÉTAPE	CE QUE CELA IMPLIQUE POUR LA PRÉCISION ET LE RISQUE
—	**Dans ce modèle, l'IA est ouverte dès le départ et les réponses sont encadrées par des étapes de contrôle, mais les hallucinations ne sont pas éliminées.**	—

Architecture COMPAiSS à exécution conditionnelle (Execution-Gated)

ÉTAPE	CE QUI SE PRODUIT À CETTE ÉTAPE	CE QUE CELA IMPLIQUE POUR LA PRÉCISION ET LE RISQUE
Inférence du modèle (Toujours ouverte)	Le modèle d'IA n'est pas actif par défaut. L'inférence est conditionnelle et ne se produit que si l'autorisation est accordée.	Le système ne suppose pas qu'il répondra avant d'avoir déterminé s'il doit le faire.
Réception de la question de l'utilisateur	La question de l'utilisateur est reçue dans n'importe quelle langue et immédiatement évaluée quant à sa portée institutionnelle et à son autorisation.	Les questions hors du champ d'autorité institutionnelle sont bloquées avant tout raisonnement par IA.
Traduction / Normalisation linguistique	Si la question n'est pas en anglais, elle est traduite uniquement afin d'en normaliser le sens. Il s'agit d'une étape linguistique seulement, et non d'une étape de recherche ou de raisonnement.	La traduction n'influence pas les sources considérées ni la détermination des réponses.
Génération d'embeddings	La question normalisée est convertie en représentation mathématique, mais uniquement à l'intérieur d'un espace de vérité institutionnel pré-délimité.	La recherche est restreinte aux sources approuvées par l'institution (liste verte) avant tout raisonnement par IA.
Récupération vectorielle / de documents (RAG)	Le système récupère uniquement à partir de sources institutionnelles autorisées. Si aucun contenu faisant autorité n'existe, le processus s'arrête ici.	Seul un contenu vérifié et approuvé par l'institution peut informer une réponse.
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 confinée aux sources institutionnelles et ne peut combler des lacunes à partir de données d'entraînement générales.

ÉTAPE	CE QUI SE PRODUIT À CETTE ÉTAPE	CE QUE CELA IMPLIQUE POUR LA PRÉCISION ET LE RISQUE
Contrôles post-génération	Étant donné que la réponse repose exclusivement sur des sources autorisées, aucun filtrage post-génération n'est nécessaire pour assurer la précision ou l'autorité.	Les hallucinations fondées sur des sources non autorisées sont structurellement empêchées, et non détectées rétroactivement.
Réponse finale livrée	La réponse est retraduite (au besoin) et présentée à l'utilisateur, fondée exclusivement sur des documents institutionnels. Si aucune réponse n'est autorisée, l'utilisateur en est informé.	Les utilisateurs reçoivent uniquement des réponses précises et autorisées, ou une indication claire qu'aucune réponse n'est disponible.
—	**Dans ce modèle, le raisonnement par IA est conditionné avant même de commencer, empêchant la génération de réponses non étayées.**	—