La voie moins empruntée - Repenser les coûts, la sécurité et la confiance liés à l'IA générative dans les institutions réglementées

Pourquoi cette conversation importe

On dit aux institutions que l'intelligence artificielle est inévitable. Pas simplement utile - inévitable. Le message est constant chez les fournisseurs, qui réitèrent généralement le même point : l'IA transformera la manière dont les organisations fonctionnent, mais il faudra mettre en place les bons garde-fous.

Ce qui est rarement remis en question, c'est de savoir si l'on demande aux institutions de s'adapter à un modèle industriel qui n'a jamais été conçu pour elles au départ. L'objectif ici n'est ni de critiquer les outils et produits d'IA standards, ni de soutenir qu'une approche serait universellement supérieure. Il s'agit plutôt d'expliquer pourquoi les organisations réglementées peuvent avoir besoin d'un type d'IA fondamentalement différent - et pourquoi cette différence compte davantage que les fonctionnalités, les modèles, les discours marketing ou le prestige des fournisseurs.

Pour clarifier le problème, j'utiliserai deux métaphores simples que j'explore tout au long de la phase de développement de COMPAiSS, un assistant d'IA conçu pour des environnements institutionnels :

- **La matrice de l'IA** - l'ensemble des hypothèses implicites qui façonnent la manière dont la plupart des systèmes d'IA sont conçus.
- **La route (et la route moins fréquentée)** - la façon dont ces hypothèses se traduisent en coûts, en risques, en mesures correctives recommandées, en arguments marketing et en adéquation institutionnelle.

Les hypothèses de la matrice de l'IA que tout le monde accepte (souvent sans s'en rendre compte)

La plupart des systèmes d'IA modernes sont construits à l'intérieur de ce que l'on peut raisonnablement appeler un état d'esprit dominant unique. Dans cet état d'esprit, on suppose que l'IA « pense » en permanence et qu'elle est toujours active. Elle est autorisée à raisonner sur le monde de tout - Internet, les connaissances générales, les faits implicites, les analogies et les conjectures. La sécurité est quelque chose que l'on gère après le début du raisonnement, et tous les risques correspondants sont jugés acceptables tant qu'ils peuvent être filtrés, modérés ou refusés.

Cet état d'esprit n'est pas malveillant ; il constitue le principe fondateur accepté de l'IA grand public, de l'IA créative et de l'IA à usage général. Et il est cohérent si l'objectif est de vendre, de retenir et de renforcer une intelligence ouverte et sans limites. Une fois pleinement intégré à cette matrice, certaines choses deviennent alors indiscutables :

- bien sûr que l'IA doit d'abord penser ;

- bien sûr que des erreurs surviendront ;
- bien sûr que les hallucinations sont inévitables ;
- bien sûr qu'il faut des couches de contrôle pour gérer les risques ;
- bien sûr que les coûts augmentent avec la complexité.

À l'intérieur du paradigme standard de l'IA, tout cela paraît évident et rassurant. Ce qui est rarement posé comme question, c'est : et si l'ensemble de ce cadre était inapproprié pour des organisations réglementées ayant l'obligation de fournir une information exacte et faisant autorité aux personnes qu'elles servent ?

La route principale : comment l'IA d'entreprise est construite (et pourquoi)

La plupart des produits d'IA d'entreprise - y compris ceux explicitement destinés à des environnements réglementés - empruntent la route principale, qui se présente ainsi :

Commencer par une IA à usage général capable de répondre à presque tout ; la connecter à des documents internes à l'aide de techniques comme la génération augmentée par récupération (« RAG ») ; ajouter des couches de modération, de filtres de politiques, d'invites de refus, de tableaux de bord de conformité, etc. ; puis espérer que, la plupart du temps, le système se comportera correctement.

Pour prolonger la métaphore de la route, cela revient à donner à tout le monde accès à un réseau autoroutier mondial, puis à investir des ressources considérables dans la gestion de la circulation, les feux de signalisation, les limites de vitesse, les caméras, la police, les rapports d'accident, les assurances et l'application continue des règles. La route est immense, les destinations sont imprévisibles, de sorte que l'application des règles ne s'arrête jamais - ce qui explique pourquoi ces systèmes sont extrêmement coûteux à déployer, coûteux à maintenir et jamais pleinement dignes de confiance, même aux yeux de leurs propres fournisseurs.

Pourquoi le RAG paraît cohérent dans la matrice - et pourquoi il laisse néanmoins passer le risque

La génération augmentée par récupération (RAG) est rassurante en apparence. L'idée est simple : si l'IA ne consulte que des documents approuvés, les hallucinations peuvent être contrôlées. Mais voici l'élément facile à manquer lorsqu'on est entièrement immergé dans le paradigme accepté : le RAG contrôle ce que l'IA consulte - et non si, ou comment, l'IA est autorisée à penser en premier lieu.

Même avec des documents parfaits, l'IA continue de raisonner, de supposer ce qui manque, de tirer des conclusions qui n'ont jamais été explicitement formulées, de deviner comment les éléments s'articulent entre eux et de relier des points qui peuvent ne pas exister.

Le RAG restreint les intrants, mais ne modifie pas la nature du raisonnement. C'est pourquoi les hallucinations continuent de se produire, simplement sous des formes plus subtiles et plus

dangereuses - avec un ton faisant autorité, en mêlant des politiques à des inférences, en comblant avec assurance ce qui n'a jamais été écrit. Pour les organisations réglementées, ce sont les pires types d'erreurs : non pas des absurdités flagrantes, mais des réponses presque justes qui sonnent comme officielles.

Le problème caché des coûts : pourquoi la gestion du risque est si onéreuse

Cela explique pourquoi les solutions d'IA d'entreprise deviennent si coûteuses avec le temps. Elles exigent davantage de surveillance, plus d'outils de conformité, plus de révision humaine, plus de mises à jour de politiques et plus de gestion des exceptions.

Ces coûts ne sont pas accidentels. Ils constituent des conséquences structurelles de la route principale, et de puissantes incitations économiques contribuent à maintenir ces structures : les fournisseurs sont récompensés pour la gestion du risque, non pour son élimination ; la complexité justifie la tarification d'entreprise ; et les couches de gouvernance deviennent des fonctionnalités facturables. De l'intérieur du paradigme de l'IA généralement accepté, cela ressemble à un progrès. Vu de l'extérieur, cela ressemble à une course aux armements contre un problème qui a été accepté comme inévitable.

La route moins fréquentée : un monde de vérité plus restreint et plus sûr

En sortant de cette matrice, la voie alternative commence par une question très différente : et si le raisonnement lui-même était conditionnel ? Non pas : « Comment nettoyer les résultats ? » ou « Comment détecter les hallucinations ? » ou « Comment repérer et refuser les mauvaises réponses ? », mais plutôt : le système devrait-il être autorisé à penser tout court - à moins d'être déjà à l'intérieur d'un monde digne de confiance ?

COMPAiSS repose sur une idée simple mais radicale : établir d'abord le monde, puis libérer l'IA pour qu'elle ne fonctionne qu'à l'intérieur de ce monde. Ce cadre est né de la confrontation répétée aux mêmes contraintes institutionnelles - la confiance, l'exactitude et la responsabilité.

Au lieu du monde de tout, le système opère à l'intérieur d'un monde de vérité préautorisé, défini par l'institution elle-même. Dans ce monde, seules les sources officielles existent, seuls les domaines approuvés sont accessibles et seule la connaissance institutionnelle est présente. L'IA n'a pas besoin d'être empêchée d'aller ailleurs - l'ailleurs n'existe pas.

En reprenant la métaphore de la route, COMPAiSS est conçu comme un système de transport construit sur mesure, où tout le monde conduit le même type de véhicule sur la même route. Imaginez maintenant une seule route bien conçue, où chaque véhicule autonome roule exactement à la limite permise, où les voies sont uniformes, les intersections rares et où tout le monde se dirige vers la même ville - même si chacun y cherche des choses différentes une fois arrivé. Dans un tel environnement, il n'est pas nécessaire d'avoir des policiers de la circulation à chaque coin de rue, des systèmes d'application complexes ou une surveillance constante des

comportements dangereux. La structure même de la route accomplit l'essentiel du travail. C'est ce que COMPAiSS fait pour le raisonnement de l'IA.

En contraignant le monde dans lequel l'IA opère, COMPAiSS élimine le besoin de nombreuses mesures de contrôle très coûteuses sur lesquelles reposent les systèmes d'IA d'entreprise. Il n'est plus nécessaire de « tirer le volant » en permanence pour éviter des directions dangereuses, puisque ces directions n'ont jamais été construites.

Chacun continue de poser des questions différentes et de rechercher des réponses détaillées, exactes et de grande qualité - mais tous le font à l'intérieur d'un environnement partagé et digne de confiance, régi par la compréhension institutionnelle de ce qui est vrai et pertinent. C'est pourquoi COMPAiSS paraît plus simple, et pourquoi son exploitation est beaucoup moins coûteuse.

Pourquoi cette approche gère les hallucinations à un niveau structurel

Les hallucinations ne résultent pas de la malveillance ; elles émergent naturellement lorsqu'un système bénéficie d'une grande liberté de raisonnement. Le modèle standard produit donc deux types d'erreurs : des faux positifs, où quelque chose d'unsafe ou d'inexact passe à travers, et des faux négatifs, où une réponse légitime est bloquée par excès de prudence. Les deux sont inévitables lorsque le système raisonne toujours en premier et est corrigé ensuite.

Avec COMPAiSS, des catégories entières d'hallucinations disparaissent - non pas parce que l'IA se comporte mieux, mais parce qu'il n'y a rien sur quoi halluciner. Le raisonnement n'est permis qu'à l'intérieur d'un monde de vérité préautorisé et défini par l'institution. Si une réponse ne peut y être ancrée, le système n'entre jamais dans un état où la conjecture est permise, éliminant ainsi la classe la plus dangereuse de réponses confiantes mais presque justes.

Bien entendu, COMPAiSS peut encore commettre des erreurs en raison d'ambiguïtés dans les sources, de politiques peu claires ou de liens manquants. Mais ce sont des erreurs plus sûres. Elles sont circonscrites, explicables et ancrées institutionnellement. En contrôlant l'endroit même où le raisonnement est autorisé à exister - plutôt qu'en tentant de surveiller chaque sortie après coup - COMPAiSS réduit à la fois la fréquence et la gravité des erreurs.

Pourquoi cette architecture est déterminante pour les secteurs réglementés

L'argument architectural présenté ici dépasse naturellement le cadre d'une seule organisation. Les universités, les hôpitaux et réseaux de soins de santé, ainsi que les services gouvernementaux ou publics, opèrent tous sous des pressions similaires : ils sont responsables de l'exactitude de l'information qu'ils fournissent, doivent préserver la confiance institutionnelle et sont de plus en plus exposés aux risques de traduction et d'interprétation à mesure que les services s'étendent à de nouvelles langues et plateformes.

Dans ces environnements, des réponses « presque justes » sont souvent plus dangereuses que des erreurs manifestes. Un léger glissement interprétatif - introduit lors de la récupération, de la traduction ou de la génération - peut entraîner des conséquences juridiques, cliniques ou politiques qui ne peuvent pas être facilement corrigées après coup.

Parallèlement, ces institutions subissent des pressions croissantes en matière de coûts et de gouvernance. Les systèmes qui nécessitent une surveillance, une remédiation et une supervision continues pour demeurer sécuritaires introduisent des charges opérationnelles à long terme qui s'accumulent avec le temps. Une architecture qui réduit le risque de façon structurelle, plutôt que de le gérer indéfiniment, s'aligne davantage sur la manière dont les organisations réglementées sont censées fonctionner et rendre des comptes.

Pourquoi cette route n'est pas pour tout le monde - et pourquoi c'est acceptable

Le modèle COMPAiSS n'est pas conçu pour répondre à tout. Il est conçu pour répondre aux bonnes choses. Les organisations réglementées n'ont pas besoin d'une IA capable d'expliquer l'univers ; elles ont besoin d'une IA capable d'expliquer leurs règles, leurs politiques, leurs services, leurs obligations et leurs décisions.

COMPAiSS détermine ce qu'il est permis de penser avant même que le raisonnement ne commence. Cette différence modifie les coûts, la sécurité, la confiance et l'adéquation institutionnelle. La route moins fréquentée n'est pas spectaculaire, mais elle est plus silencieuse, moins coûteuse, plus sûre et plus honnête. Et pour des institutions fondées sur la confiance, c'est précisément l'essentiel.