COMPAiSS

Sortir de la matrice

Une nouvelle opportunité pour les cabinets-conseils en IA avec des clients d'industries réglementées

Frank P. Harvey, PhD
Conseiller principal, Université Dalhousie
Fondateur, chef de la direction et architecte en chef, **COMPAiSS Inc.**

Mai 2026

Brevet en instance : OPIC 3 299 174 (Canada) / USPTO 19/455 963 (États-Unis)
Contact : frank.harvey@dal.ca

Résumé

Ce livre blanc s'adresse aux principaux cabinets-conseils qui accompagnent les institutions réglementées en matière de gouvernance de l'IA. Il documente une évaluation empirique multi-modèle confirmant que les grands cabinets-conseils opèrent dans le cadre d'un paradigme génération-d'abord, où les hallucinations sont considérées comme inévitables et gérées après le processus d'inférence. Il fait valoir que COMPAiSS introduit un principe architectural alternatif, l'inférence à exécution contrôlée, qui supprime les conditions structurelles favorisant les hallucinations hors portée. Il ne prétend pas offrir un système parfait. Il formule une affirmation plus précise : les erreurs qu'un tel système produit sont qualitativement différentes de celles des systèmes non contrôlés, et ce d'une manière que les institutions réglementées peuvent gouverner, vérifier et assumer. Des déploiements actifs à Service Canada, à l'Université Dalhousie et à l'Université McGill fournissent une validation empirique. L'architecture fait l'objet de demandes de brevet en cours d'examen au Canada et aux États-Unis.

1. Le paradigme dominant

1.1 Pourquoi le paradigme génération-d'abord persiste

Le paradigme génération-d'abord est le fruit de l'évolution de l'industrie de l'IA. Les grands modèles de langage ont été conçus pour offrir une grande polyvalence : capacité à raisonner dans de nombreux domaines, à inférer le contexte et à produire des réponses fluides sur presque

n'importe quel sujet. Pour les applications grand public, la génération de contenu, la productivité interne et la recherche tous azimuts, cette capacité est génuinement précieuse.

Le conseil en IA aux entreprises s'est développé à l'intérieur de ce paradigme. L'approche consultative standard consiste à connecter un modèle polyvalent aux données d'une organisation par le biais de la génération augmentée par récupération (GAR), à ajouter des couches de gouvernance, à créer des tableaux de bord de conformité et à mettre en place des flux d'examen humain. Il s'agit d'une réponse cohérente et justifiable à un défi réel, et les pratiques qui en découlent témoignent d'un investissement et d'une expertise considérables.

Le paradigme génération-d'abord obéit également à une logique économique structurelle. Puisque l'inférence est toujours autorisée à s'exécuter, les institutions doivent continuellement investir dans l'infrastructure compensatoire qui gère ses sorties. Les outils de modération, les couches de sécurité, les systèmes d'audit et les flux d'examen humain sont de véritables exigences opérationnelles, non des complexités inventées. Ce sont les coûts nécessaires de la gestion du risque dans un système où la réflexion s'effectue toujours en premier. Les cabinets-conseils qui ont bâti leurs pratiques autour de cette infrastructure l'ont fait parce qu'elle était, et demeure, nécessaire pour les déploiements génération-d'abord.

Ce livre blanc ne soutient pas que ces pratiques ont mal été conçues. Il soutient que, pour une catégorie spécifique d'institutions, dans un contexte de déploiement précis, un point de départ architectural différent est disponible, et que les cabinets-conseils sont bien placés pour guider leurs clients vers cette alternative.

1.2 Ce que quatre modèles d'IA indépendants ont révélé

COMPAiSS a mené une évaluation multi-modèle des recommandations de gouvernance de l'IA diffusées publiquement par les principaux cabinets-conseils aux institutions réglementées. Quatre grands modèles de langage indépendants, GPT, Gemini, Copilot et Perplexity, ont reçu chacun une invite identique centrée sur la façon dont ces cabinets conceptualisent et contrôlent les hallucinations de l'IA. Chaque modèle a produit une évaluation autonome. Les quatre évaluations ont ensuite été examinées pour en dégager le consensus et les divergences.

La conception méthodologique mérite d'être explicitée. Chaque modèle a reçu la même invite strictement délimitée et a été limité à la documentation produit, aux cadres de gouvernance et aux descriptions architecturales accessibles au public. Aucun modèle n'a reçu les sorties des autres durant la phase d'évaluation. Le consensus reflète donc une convergence indépendante, non un raisonnement circulaire.

Constat de consensus multi-modèle (GPT, Gemini, Copilot, Perplexity)

Les quatre modèles ont conclu indépendamment que les principaux cabinets-conseils opèrent dans le cadre d'un paradigme génération-d'abord avec atténuation post-hoc. Les hallucinations sont traitées comme des propriétés inhérentes et probabilistes des grands modèles de langage, à gérer par la gouvernance, la GAR et la supervision humaine après que l'inférence a déjà eu lieu.

Aucun cabinet n'a été identifié comme prônant le refus d'exécution pré-inférence ou la prévention structurelle des hallucinations. Dans chaque cadre examiné, le refus est lui-même une sortie de l'exécution générative. Le modèle a fonctionné et a produit un refus. C'est architecturalement distinct d'un système dans lequel le modèle n'a jamais fonctionné.

Cabinet	Ce que le cadre traite	Ce que l'évaluation multi-modèle a révélé
Deloitte	Cadres de gouvernance, assurance qualité et révision humaine après génération	Les hallucinations sont traitées comme un risque inévitable de fiabilité ; aucun mécanisme de prévention structurelle identifié
PwC	GAR et discipline d'invite pour réduire les sorties plausibles mais sans fondement	La génération se produit toujours en premier ; l'atténuation est appliquée pendant et après, non avant
KPMG	Auditabilité, surveillance et contrôles de gouvernance après génération	Les hallucinations sont traitées comme endémiques ; abordées de manière réactive par des cadres de surveillance
Microsoft	Ancrage, boucles de correction et outils de sécurité (Bedrock Guardrails, Azure AI Content Safety)	Les hallucinations définies comme contenu non ancré ; réduites, non prévenues ; l'inférence s'exécute toujours

> *GPT ne trouve aucune preuve de refus d'exécution pré-inférence ni de prévention structurelle au sein d'aucun cabinet. Gemini énonce explicitement qu'aucun des quatre cabinets ne remet en question l'hypothèse génération-d'abord. Copilot ne trouve aucune préconisation de blocage pré-génération ; le refus est traité comme un comportement configuré, non comme une architecture par défaut. Perplexity note explicitement que la GAR est traitée comme un mécanisme compensatoire, non comme une porte à vérou dur empêchant l'inférence. Source : COMPAiSS, Hallucinations par conception : une évaluation multi-modèle, 2026*

1.3 Le problème structurel des hallucinations

La persistance des hallucinations dans toutes les approches d'atténuation actuelles n'est pas un échec d'efforts d'ingénierie. C'est la conséquence d'une hypothèse architecturale partagée : l'inférence générative s'exécute en premier, et la sécurité est appliquée après. Chaque système GAR, chaque garde-fou, chaque couche de modération, chaque flux d'examen humain opère sur des sorties que le modèle génératif a déjà produites.

La GAR est la recommandation consultative dominante pour les institutions réglementées. L'intuition qui la sous-tend est juste : si l'IA ne consulte que des documents approuvés, les hallucinations peuvent être contrôlées. La limite réside dans ce que la GAR fait réellement par rapport à ce dont les institutions ont besoin. La GAR contrôle ce que l'IA lit, pas si ni comment l'IA est autorisée à raisonner. Même avec un corpus de récupération bien sélectionné, le modèle génératif sous-jacent conserve la capacité d'extrapoler à partir de sa mémoire paramétrique, d'interpréter incorrectement le contenu récupéré et de produire des sorties plausibles non ancrées dans ce qui a été récupéré.

Taux d'hallucination documentés dans les systèmes d'IA d'entreprise (2023 à 2025)

- 17 à 34 % : taux d'hallucination pour Westlaw AI-Assisted Research et Lexis+ AI sur des requêtes juridiques réelles. Ces plateformes GAR spécialisées sont commercialisées comme « sans hallucination » par Thomson Reuters et LexisNexis. (Magesh et al., Stanford RegLab / HAI, 2024)

- 50 à 83 % : taux d'hallucination adversariale sur six grands modèles d'IA lors de tests de vignettes cliniques validées par des médecins. (Omar et al., Nature Medicine / Communications Medicine, 2025)
- 33 à 51 % : taux d'hallucination pour les modèles de raisonnement OpenAI o3 sur des repères factuels domaine ouvert, plus élevés que les modèles de génération précédente malgré une capacité accrue. (Graffius, 2026)
- 6 % : taux résiduel d'hallucination dans des conditions GAR optimales pour des tâches de résumé ancré très contraintes. C'est le scénario du meilleur cas dans des conditions contrôlées. (Nishisako, Higashi et Wakao, 2025)

Le constat central des recherches de 2023 à 2025 : l'hallucination n'est pas une limitation temporaire que l'amélioration des modèles éliminera. C'est une propriété structurelle persistante des architectures inférence-d'abord. Les modèles plus capables n'hallucinenent pas automatiquement moins. Sur des tâches de raisonnement complexes, certains hallucinenent davantage.

2. L'alternative architecturale

2.1 L'inférence à exécution contrôlée

COMPAiSS introduit un principe architectural différent. Plutôt que de générer une réponse puis de la filtrer, COMPAiSS évalue l'autorisation avant que toute inférence générative ne soit autorisée à se produire.

Comparaison architecturale

Génération-d'abord (tous les systèmes d'entreprise actuels) :

Requête reçue > Exécution d'inférence instanciée > Modèle génératif exécuté > Sortie produite > Sécurité / filtrage / modération appliqués > Réponse livrée

COMPAiSS : inférence à exécution contrôlée :

Requête reçue > Porte d'autorisation évalue les sources institutionnelles > SI autorisé : exécution d'inférence instanciée > Réponse générée à partir des sources approuvées > Réponse livrée

SI non autorisé : aucune exécution d'inférence instanciée. Aucun calcul génératif exécuté. Réponse d'échec sécurisé livrée à coût marginal zéro.

La porte d'exécution pré-inférence n'est pas un garde-fou, un filtre ni une couche de modération appliquée à un service d'inférence en cours d'exécution. C'est une condition qui détermine si une exécution d'inférence peut exister pour une requête donnée. Lorsque l'autorisation échoue parce qu'aucune source institutionnelle n'appuie une réponse, aucun calcul génératif ne se produit. Il n'y a pas d'inférence non autorisée d'où des hallucinations hors portée pourraient surgir.

Cette distinction n'est pas sémantique. Dans les systèmes génération-d'abord : *le modèle existe et l'autorisation régit ce qu'il produit.* Dans les systèmes à exécution contrôlée : *l'autorisation*

détermine si le modèle existe. Le modèle ne s'exécute que lorsque les sources autorisées le confirment.

Dans la portée autorisée, COMPAiSS emploie une architecture de défense en profondeur. Des contraintes basées sur des instructions guident le comportement du modèle lors de l'inférence autorisée. La validation post-génération des URL s'assure que tous les liens pointent vers des sources institutionnelles autorisées. Ces contrôles réduisent le risque de génération dans la portée autorisée sans l'éliminer. La section 2.2 aborde cela directement.

2.2 Ce que COMPAiSS ne prétend pas

Certains réviseurs d'une version antérieure de ce document ont noté que certaines formulations laissaient entendre que le contrôle d'exécution élimine les hallucinations de façon générale. Ce n'est pas l'affirmation, et la distinction importe pour un public de gouvernance.

COMPAiSS formule une affirmation précise, tirée de sa propre documentation technique :

> *« COMPAiSS élimine par conception les hallucinations liées aux violations de portée, et réduit sensiblement, sans les éliminer, les risques de qualité de génération dans la portée autorisée grâce à l'analyse structurée et à des contextes d'inférence étroitement délimités. » Source : Documentation technique COMPAiSS*

Dans la portée autorisée, le modèle génératif sous-jacent peut encore mal interpréter des éléments de preuve, agréger incorrectement de l'information provenant de plusieurs sources ou produire des réponses techniquement ancrées mais pratiquement trompeuses. Ces risques résiduels sont réels et reconnus. Ils sont également qualitativement différents des risques d'hallucination dans les systèmes non contrôlés, et cette différence détermine si les institutions réglementées peuvent gouverner et assumer leurs déploiements d'IA.

La question pertinente n'est jamais de savoir si un système est parfait. Aucun système d'IA ne l'est. La question pertinente est : quels types d'erreurs ce système produit-il, et l'institution peut-elle les détecter, les gouverner et les assumer ? Le tableau ci-dessous répond directement à cette question.

2.3 Comparaison des profils de défaillance

Le bien-fondé de l'inférence à exécution contrôlée ne repose pas sur l'affirmation que COMPAiSS ne fait jamais d'erreurs. Il repose sur l'affirmation que les erreurs qu'il commet sont qualitativement différentes de celles des systèmes non contrôlés, d'une manière qui importe spécifiquement pour la gouvernance institutionnelle réglementée.

Dimension	COMPAiSS : erreurs dans la portée autorisée (résiduelles)	IA génération-d'abord : risque d'hallucination (structurel)
Classe d'erreur	Navigationnelle ou interprétative : le client peut être dirigé vers une page institutionnelle autorisée moins pertinente	Fabricatoire ou décisionnelle : règle d'admissibilité, montant de prestation, date limite ou détail de politique halluciné, présenté avec pleine assurance
Ancrage dans les sources	Chaque réponse renvoie à au moins une URL de source institutionnelle vérifiée et approuvée	La réponse peut ne citer aucune source, une source fabriquée ou une source réelle qui n'étaye pas l'affirmation faite
Autorité décisionnelle	Aucune : le système fournit de l'information seulement et ne détermine	Aucune formellement, mais les réponses erronées empreintes

	pas l'admissibilité, ne calcule pas les prestations ni ne statue sur les réclamations	d'assurance sont couramment utilisées comme si elles faisaient autorité
Rectifiabilité	Immédiatement rectifiable : le client peut reformuler sa requête, suivre le lien source ou contacter le personnel	Souvent non reconnue comme incorrecte par l'utilisateur ; l'erreur peut ne se révéler qu'après qu'une décision conséquente ait été prise
Fréquence avec la GAR (entreprises)	Erreurs navigationnelles dans la portée autorisée : non repérées indépendamment ; le journal d'audit permet la détection et la correction	17 à 34 % sur les requêtes réelles dans des plateformes juridiques GAR spécialisées (Magesh et al., Stanford/Yale, 2025)
Conséquence dans le pire cas	L'utilisateur contacte le personnel pour clarification ; aucun changement à ses droits ou prestations	Un citoyen vulnérable agit sur la foi d'un montant de prestation halluciné, d'une date limite manquée ou d'une règle d'admissibilité fausse, aux conséquences potentiellement irréversibles

Remarque : les erreurs navigationnelles dans la portée autorisée de COMPAiSS ne sont pas repérées indépendamment sur un ensemble de données public. Le journal d'audit disponible pour les administrateurs institutionnels permet la détection et la correction. Les taux d'hallucination génération-d'abord cités reflètent des recherches publiées, révisées par des pairs, portant sur des systèmes d'entreprise dans des domaines réglementés.

L'asymétrie du tableau ci-dessus est l'argument de gouvernance. Les erreurs que COMPAiSS produit sont délimitées, traçables jusqu'à leur source, rectifiables et n'ont aucune autorité décisionnelle. Les erreurs que les systèmes génération-d'abord produisent dans les environnements réglementés sont documentées à un taux de 17 à 34 % sur des plateformes d'entreprise spécialisées, structurellement inévitables, et dans le pire des cas, elles sont exploitées par des personnes vulnérables qui n'ont aucun moyen de savoir que l'information qu'elles ont reçue a été fabriquée.

> *« Le pire résultat d'une réponse incorrecte de COMPAiSS est navigationnel ou interprétatif, non fabricatoire ni décisionnel : le client peut être dirigé vers une page autorisée moins pertinente. Le pire résultat d'une réponse IA non contrôlée est une réponse fausse, spécifique et assurée, un montant de prestation halluciné, une règle d'admissibilité ou une date limite erronée, qu'un client vulnérable exploite sans savoir que l'information était fausse. » Source : Livre blanc COMPAiSS, Inférence à exécution contrôlée, 2026*

2.4 Deux points d'entrée

COMPAiSS opère à deux points d'entrée naturels dans le parcours d'IA d'une institution.

Pour les institutions ayant déjà investi dans la GAR : COMPAiSS opère comme une couche de gouvernance pré-inférence qui résout les taux d'hallucination résiduels que la GAR seule ne peut régler. Il ne s'agit pas d'une recommandation de remplacement de l'infrastructure existante. C'est un contrôle de gouvernance qui comble la lacune structurelle que l'architecture génération-d'abord laisse ouverte. Les cabinets-conseils peuvent le présenter comme une mise à niveau de la maturité de gouvernance, aidant les institutions ayant investi dans l'infrastructure GAR à atteindre le seuil de responsabilité que les environnements réglementés exigent.

Pour les institutions sans investissement GAR : qu'elles soient limitées par les coûts ou préoccupées par les taux de défaillance documentés dans les meilleures conditions GAR, ces institutions disposent d'une alternative directe. COMPAiSS fournit une architecture d'IA

entièrement gouvernée qui fait respecter la portée, l'autorité des sources et la politique institutionnelle sans nécessiter d'investissement en infrastructure d'entreprise. Pour ces clients, les cabinets-conseils peuvent présenter COMPAiSS comme le parcours gouvernance-d'abord.

2.5 Structure de coûts

Le bien-fondé financier découle directement de l'architecture. Parce qu'environ 40 % des requêtes se voient refuser l'inférence à coût marginal zéro, et parce que les requêtes autorisées n'insèrent pas de documents récupérés dans le contexte du modèle, COMPAiSS consomme sensiblement moins de ressources de calcul que les déploiements GAR conventionnels.

Composante de coût	GAR conventionnelle	COMPAiSS
Inférence (jetons)	10 000 $ à 30 000 $	7 000 $ à 18 000 $
Infrastructure GAR	30 000 $ à 60 000 $	0 $
Outils de sécurité et de modération	30 000 $ à 75 000 $	0 $
Surveillance et conformité	10 000 $ à 20 000 $	5 000 $ à 15 000 $
COÛT ANNUEL TOTAL	**90 000 $ à 200 000 $**	**15 000 $ à 30 000 $**

Hypothèses : environ 100 000 requêtes annuelles, taux d'autorisation d'environ 60 %. Les chiffres GAR conventionnels reflètent les coûts d'exploitation annuels totaux incluant le calcul d'inférence, l'infrastructure de base de données vectorielles persistante et les outils de gouvernance compensatoires, basés sur des repères de déploiement GAR d'entreprise publiés. Les chiffres COMPAiSS reflètent les coûts d'API d'inférence aux tarifs actuels des modèles, la récupération par liste verte et l'hébergement de la plateforme. Les coûts réels varient selon le volume de requêtes, la sélection du modèle et la configuration infonuagique. Ces chiffres sont indicatifs et directionnels, non contractuels.

L'écart de coût de 6 à 12 fois n'est pas obtenu par une capacité réduite. C'est une conséquence structurelle de l'architecture à exécution contrôlée : le coût est évité parce que les requêtes non autorisées ne génèrent aucune sortie IA du tout.

2.6 Déployé et validé

Déploiements institutionnels actifs et tests bêta (mai 2026)

Service Canada (gouvernement fédéral)

Assistant d'information IA couvrant l'assurance-emploi, le Régime de pensions du Canada, la Sécurité de la vieillesse, le Supplément de revenu garanti et les services liés au numéro d'assurance sociale. Entièrement bilingue avec acheminement aux sources en français confirmé. Prise en charge multilingue des requêtes testée, notamment en tagalog, avec récupération de taux de prestations de l'année en cours exacts à partir des sources officielles du gouvernement du Canada. Audit d'accessibilité WCAG 2.1 AA — Phase 1 complétée ; phase 2 en cours.

Université Dalhousie

Requêtes sur les affaires étudiantes et les politiques académiques couvrant le bien-être étudiant, le statut académique, les appels de notes et les procédures d'accommodement. Dalhousie a mené une évaluation institutionnelle formelle et a sélectionné COMPAiSS à la

suite d'une évaluation concurrentielle incluant un fournisseur d'IA génération-d'abord bien établi dans l'enseignement supérieur.

Université McGill

Requêtes sur les admissions au premier cycle et le statut académique. Les tests de frontière ont confirmé que le système cite les sources avec précision lorsque les sources autorisées contiennent une réponse, et reconnaît explicitement les lacunes dans le cas contraire, dirigeant les étudiants vers le bureau concerné plutôt que de produire une réponse fabriquée plausible. Aucune date n'a été générée là où aucune date n'existait dans la source autorisée. Actuellement en phase de test bêta.

3. L'avantage pour les cabinets-conseils

3.1 L'opportunité consultative

Les institutions réglementées subissent une pression de gouvernance croissante de plusieurs côtés simultanément. Le GAO américain a documenté que les cas d'utilisation de l'IA par les agences fédérales ont multiplié par neuf entre 2023 et 2024, avec plus de 85 % des cas d'utilisation à fort impact sans la documentation de gestion des risques requise. Des tribunaux français ont émis des avertissements judiciaires formels concernant le contenu juridique halluciné par l'IA atteignant les procédures judiciaires fin 2025. Le gouvernement du Québec a publié des orientations réglementaires formelles mettant en garde contre les conséquences pour les citoyens qui s'appuient sur l'information générée par l'IA pour des décisions financières ou de santé. Le Cadre de gestion des risques liés à l'IA du NIST exige la documentation du risque résiduel, reconnaissant que tous les incidents et défaillances ne peuvent être éliminés dans les architectures actuelles.

Les institutions qui font face à cette pression ont besoin de conseillers capables d'offrir plus qu'une couche de gouvernance supplémentaire appliquée à un système génération-d'abord. Le cabinet-conseil capable d'expliquer l'inférence à exécution contrôlée au conseil d'administration d'un hôpital, au sénat d'une université ou à un directeur général de l'information gouvernemental, et d'aider cette institution à évaluer si ses obligations de gouvernance sont compatibles avec un taux d'hallucination résiduel, offre quelque chose qualitativement différent de ce qui est présentement disponible.

3.2 Trois lignes de services consultatifs

Opportunités de services consultatifs

1. Analyse de l'architecture de gouvernance

Aider les institutions réglementées à évaluer si leur architecture IA actuelle ou planifiée répond à leurs obligations de gouvernance. La question centrale n'est pas quel produit acheter. C'est de déterminer si le cadre de responsabilité de l'institution est compatible avec un taux d'hallucination résiduel, et dans la négative, quelle alternative architecturale répond à ses obligations.

2. Mise en oeuvre et conception de la portée

Pour les institutions qui adoptent l'inférence à exécution contrôlée, le mandat de mise en oeuvre est substantiel : définition de la portée institutionnelle, architecture de la liste verte, intégration aux systèmes existants et préparation du personnel. Ce travail bénéficie d'une expérience consultative dans les environnements réglementés d'une manière qu'un fournisseur technologique seul ne peut reproduire.

3. Conseil continu en gouvernance

L'IA à exécution contrôlée exige un jugement institutionnel continu : maintenance de la liste verte à mesure que les politiques évoluent, décisions sur les frontières de portée à mesure que de nouveaux programmes s'ajoutent, examen du journal d'audit et rapports de gouvernance. Cette couche consultative est directement liée aux obligations de gouvernance de l'institution plutôt qu'à un produit spécifique.

3.3 Répondre à l'objection interne

L'objection interne franche à tout cabinet-conseil lisant ce document est directe : si l'inférence à exécution contrôlée est adoptée à grande échelle, la pratique consultative de remédiation rétrécit. C'est une observation équitable qui mérite une réponse directe.

Le modèle de remédiation fait face à ses propres vents contraires. À mesure que les échecs d'hallucination deviennent plus publics et plus lourds de conséquences juridiques, le cabinet-conseil associé à la gouvernance génération-d'abord s'expose à des risques de réputation que n'encourt pas un cabinet conseillant des alternatives gouvernance-d'abord. La question n'est pas de savoir si le paradigme évoluera éventuellement sous la pression réglementaire et institutionnelle. La question est de savoir si un cabinet mène ce changement ou y réagit après qu'un concurrent l'a initialisé.

De nombreux partenaires de cabinets pourront trouver plus commode de présenter initialement l'inférence à exécution contrôlée comme une architecture de contrôle gouvernance-d'abord pouvant s'associer ou se placer en amont des systèmes génération-d'abord, plutôt que comme un remplacement. Ce positionnement est exact, commercialement sensé et appuyé par la structure à deux points d'entrée décrite à la section 2.4.

3.4 Le vent réglementaire favorable

- Le GAO américain a documenté que plus de 85 % des cas d'utilisation fédéraux à fort impact manquaient de la documentation de gestion des risques requise en 2024, représentant à la fois un écart de conformité et une opportunité consultative.
- La loi européenne sur l'IA classe les systèmes d'IA utilisés dans l'éducation, l'emploi, les soins de santé et les services publics essentiels comme à risque élevé, avec des exigences de supervision humaine, d'auditabilité et de transparence que les systèmes génération-d'abord avec des taux d'hallucination résiduels peinent à satisfaire.
- La Directive du Canada sur la prise de décision automatisée crée des obligations de responsabilité pour les agences fédérales qui sont le plus aisément satisfaites par des systèmes dont les sorties sont déterministes et traçables jusqu'à des sources autorisées.
- Les orientations du gouvernement du Québec sur les risques liés à l'IA mettent explicitement en garde contre les conséquences pour les citoyens qui s'appuient sur l'information générée par l'IA pour des décisions financières ou de santé.

Chacun de ces développements crée une lacune de gouvernance que les institutions réglementées ont besoin d'aide pour combler. Le cabinet-conseil capable d'offrir une réponse architecturale, plutôt qu'un cadre supplémentaire pour gérer un risque inévitable, est le mieux positionné pour s'approprier ce mandat.

3.5 Une note sur le ton

Ce document utilise les analogies de la Matrice et de la Voie moins empruntée parce qu'il s'agit du cadre conceptuel publié de l'auteur et parce qu'elles communiquent efficacement une distinction architecturale réelle. Il convient toutefois d'être explicite sur ce que ces analogies sont et ne sont pas.

Elles ne constituent pas des arguments selon lesquels les cabinets-conseils ont mal géré la gouvernance de l'IA. Le paradigme génération-d'abord est rationnel pour les cas d'utilisation auxquels il a été destiné. Elles ne constituent pas non plus des arguments selon lesquels COMPAiSS est la seule architecture de gouvernance défendable. Pour l'IA polyvalente, la recherche de connaissances en entreprise et les applications tous azimuts, les systèmes génération-d'abord servent bien leurs objectifs.

Elles soutiennent que, pour une catégorie spécifique d'institutions, dans un contexte de déploiement précis, où les citoyens s'appuient sur l'information médiatisée par l'IA pour prendre des décisions concernant leurs prestations, leur santé, leur statut juridique ou leurs droits en matière d'éducation, le point de départ architectural importe d'une manière qu'aucune remédiation en aval ne peut pleinement corriger. C'est une affirmation étroite, précise et empiriquement étayée. C'est l'affirmation que ce document formule.

Conclusion

L'intuition architecturale la plus importante de ce document n'est pas que les taux d'hallucination peuvent être réduits. C'est que l'existence même d'une exécution d'inférence générative peut être soumise à un contrôle de gouvernance. Ce glissement, de la gestion de ce qu'un modèle produit à la gouvernance du fait qu'un modèle s'exécute, est la distinction fondamentale entre l'atténuation des hallucinations et la prévention structurelle des hallucinations hors portée.

COMPAiSS ne produit pas un système parfait. Dans la portée autorisée, les risques de qualité de génération demeurent et sont gérés par des contrôles de défense en profondeur. L'affirmation honnête et précise est la suivante : le contrôle d'exécution supprime les conditions structurelles qui favorisent les hallucinations hors portée, et les erreurs qui subsistent dans la portée autorisée sont qualitativement différentes de celles des systèmes non contrôlés, d'une manière que les institutions réglementées peuvent gouverner, vérifier et assumer.

Pour les cabinets-conseils qui accompagnent les institutions réglementées, l'implication pratique est claire. Les institutions posent, avec une urgence croissante, la question de savoir si leurs obligations de gouvernance en matière d'IA sont compatibles avec un taux d'hallucination résiduel. Le cabinet capable de répondre à cette question de manière architecturale, plutôt qu'avec une couche supplémentaire de remédiation, offre quelque chose que le marché actuel ne fournit pas encore à grande échelle.

COMPAiSS est en déploiement actif dans un organisme fédéral de prestation de services, dans une université de recherche qui l'a sélectionné préférablement à un concurrent génération-d'abord bien établi, et dans une deuxième université de recherche opérant dans des services aux étudiants à enjeux élevés et des domaines d'admission. L'architecture est en cours d'examen de brevet au Canada et aux États-Unis. Les données probantes sont révisées par des pairs et publiquement documentées.

La voie moins empruntée n'est pas une critique de la voie principale. C'est le constat que, pour certaines institutions ayant des obligations de gouvernance spécifiques, une voie différente est disponible. Les cabinets-conseils capables d'expliquer cette différence et de guider les institutions pour lesquelles elle importe vers le bon choix mèneront le prochain chapitre du conseil en gouvernance de l'IA dans les industries réglementées.

Pour de plus amples renseignements ou pour discuter d'un partenariat consultatif :

Frank P. Harvey, PhD | frank.harvey@dal.ca | compaiss.ca

Brevet en instance : OPIC 3 299 174 (Canada) / USPTO 19/455 963 (États-Unis)

Références

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., et Ho, D. E. (2025). Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. Journal of Empirical Legal Studies. doi:10.1111/jels.12413

Omar, M., et al. (2025). Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Communications Medicine (Nature Publishing Group). PMCID: PMC12318031.

Graffius, S.M. (2026). Are AI Hallucinations Getting Better or Worse? We Analyzed the Data. DOI: 10.13140/RG.2.2.33179.53285.

Nishisako, T., Higashi, T., et Wakao, R. (2025). Hallucination rates in constrained RAG summarization. Cité dans le livre blanc COMPAiSS, 2026.

Sun, et al. (2025). ReDeEP: Detecting Hallucination in LLMs via Decoupled Representation of Retrieval and Encoding. ICLR 2025.

Harvey, F. (2026). COMPAiSS : architecture d'inférence à exécution contrôlée. Demandes de brevet : OPIC 3 299 174 (Canada) ; USPTO 19/455 963 (États-Unis).

National Institute of Standards and Technology. (2023). AI Risk Management Framework (AI RMF 1.0). NIST AI 100-1.

US Government Accountability Office. (2025). Artificial Intelligence: Generative AI Use and Management at Federal Agencies. GAO-25-107653.

Gouvernement du Québec. (2025). Risques liés à l'intelligence artificielle. quebec.ca.

OpenAI. (2024). Why language models hallucinate. openai.com/index/why-language-models-hallucinate/

COMPAiSS Inc. (2026). Hallucinations par conception : une évaluation multi-modèle. compaiss.ca.

COMPAiSS Inc. (2026). La voie moins empruntée : repenser les coûts, la sécurité et la confiance envers l'IA générative dans les institutions réglementées. compaiss.ca.

COMPAiSS Inc. (2026). Des économies par conception : pourquoi COMPAiSS coûte moins sans sacrifier la précision. compaiss.ca.

Secrétariat du Conseil du Trésor du Canada. (2019, modifiée en 2023). Directive sur la prise de décision automatisée. Canada.ca.