COMPAiSS

Inférence à exécution conditionnelle

Une solution architecturale aux hallucinations de l'intelligence artificielle en contexte institutionnel réglementé

Frank P. Harvey, Ph. D.
Conseiller principal, Université Dalhousie
Fondateur, PDG 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
compaiss.ca

Résumé

L'hallucination par intelligence artificielle est universellement reconnue comme une limitation structurelle des systèmes d'IA générative, et non comme un défaut accidentel. Les stratégies d'atténuation existantes - notamment la génération augmentée par récupération (RAG), le filtrage après génération, les scores de confiance et la supervision humaine - réduisent les taux d'hallucination sans pouvoir les éliminer. Des recherches empiriques évaluées par les pairs confirment que même les systèmes d'entreprise fondés sur le RAG les plus sophistiqués produisent des hallucinations entre 17 % et 33 % du temps dans des environnements réglementés à enjeux élevés. Le présent article soutient que la persistance des hallucinations, malgré toutes les approches d'atténuation actuelles, est la conséquence directe d'une hypothèse architecturale commune : que l'inférence générative est toujours instanciée et que la sécurité est appliquée après coup. COMPAiSS introduit un principe architectural alternatif - l'inférence à exécution conditionnelle - dans lequel un environnement d'inférence générative n'est instancié que lorsqu'un verrou d'exécution préalable confirme l'existence de sources institutionnelles vérifiées pour étayer une réponse. En l'absence d'autorisation, aucun calcul génératif n'est effectué. Il n'existe aucune inférence non autorisée à partir de laquelle des hallucinations pourraient survenir. Le présent article décrit l'architecture, ses fondements théoriques, ses performances empiriques, sa distinction par rapport à l'art antérieur et son application aux environnements institutionnels réglementés, notamment les services gouvernementaux, les soins de santé, les services juridiques et l'enseignement supérieur. L'architecture COMPAiSS fait l'objet de demandes de brevets actuellement en cours d'examen au Canada et aux États-Unis.

1. Le problème des hallucinations est structurel, non accidentel

Les grands modèles de langage (LLM) génèrent du texte en prédisant les suites statistiquement probables de séquences d'entrée. Ce ne sont pas des machines à vérité. Ils produisent des résultats vraisemblables calibrés pour approximer la distribution du langage humain, et non pour représenter des faits vérifiés. Cette réalité architecturale - l'entraînement par prédiction du prochain jeton sans étiquettes négatives explicites - est la cause profonde des hallucinations de l'IA.

Il ne s'agit pas d'un défaut propre à un modèle ou à un produit particulier, mais d'une propriété de la classe de systèmes. Les recherches internes d'OpenAI décrivent explicitement les hallucinations comme "un défi fondamental pour tous les grands modèles de langage", en soulignant que "la précision n'atteindra jamais 100 %" parce que certaines questions sont intrinsèquement sans réponse ou ambiguës, et que les évaluations standard fondées sur la précision "récompensent les suppositions plutôt que la reconnaissance de l'incertitude".

Le consensus scientifique est sans équivoque : les hallucinations constituent une limitation inhérente et persistante des architectures LLM actuelles. Une enquête exhaustive de 2024-2025 conclut que les méthodes de détection et d'atténuation de pointe "ne parviennent pas à éliminer complètement les hallucinations", qualifiant ce problème de "défi ouvert majeur" et le présentant comme un problème de recherche en cours, non résolu. Le cadre de gestion des risques liés à l'IA du NIST demande explicitement aux organisations de documenter le risque résiduel - le risque subsistant après l'application de toutes les mesures d'atténuation - reconnaissant que "tous les incidents et défaillances ne peuvent être éliminés".

> ***"Les outils de recherche juridique fondés sur le RAG disponibles dans le commerce produisent encore des hallucinations. Près d'une requête sur cinq a amené les outils testés à fournir des informations trompeuses ou fausses." - Magesh et al. (Stanford / Yale), Journal of Empirical Legal Studies, 2025***

Les conséquences pratiques de ce risque résiduel dépendent entièrement du contexte de déploiement. Pour les systèmes d'IA grand public à usage général, les hallucinations occasionnelles constituent un compromis acceptable au regard de la large utilité offerte. Pour les environnements institutionnels réglementés - programmes de prestations gouvernementales, soins de santé, services juridiques, enseignement supérieur - les conséquences sont d'une nature catégoriquement différente. Un citoyen qui reçoit des informations erronées sur son admissibilité à l'assurance-emploi, ses droits à la Sécurité de la vieillesse ou ses droits légaux pourrait agir sur la base de ces informations avec des conséquences graves et irréversibles.

La réponse du secteur à ce problème a consisté à investir dans l'atténuation. Le RAG, les garde-fous, le filtrage après génération, les scores de confiance, les flux de révision humaine et les systèmes de modération multicouches ont tous été développés, déployés et perfectionnés. Chaque approche réduit les taux d'hallucination. Aucune ne les élimine. Le présent article soutient que la raison en est architecturale : toutes les approches d'atténuation actuelles partagent l'hypothèse que l'inférence générative s'exécute d'abord et que la sécurité est appliquée ensuite. C'est cette hypothèse qui pose problème.

2. Les limites de l'architecture à génération prioritaire

Chaque grande approche d'atténuation des hallucinations de l'IA en déploiement actuel - RAG, garde-fous, modération, filtrage, révision humaine - opère sur des résultats déjà produits par un modèle génératif. Le modèle d'exécution standard est le suivant :

> **Cycle de vie d'exécution standard de l'IA (tous les systèmes actuels)**
> Requête reçue → Environnement d'inférence instancié → Modèle génératif exécuté → Résultat produit → Sécurité / filtrage / modération appliqués → Réponse transmise

Dans ce paradigme, les mécanismes d'autorisation et de sécurité sont des contrôles d'admission et des filtres post-génération. Ils régissent l'accès à un service d'inférence déjà instancié, ou évaluent les résultats après l'exécution du modèle. Le calcul génératif - le processus stochastique susceptible de produire des hallucinations - s'est déjà produit, indépendamment du fait qu'un résultat particulier parvienne ou non à l'utilisateur.

Il ne s'agit pas d'une lacune. Cela reflète une décision architecturale délibérée et rationnelle pour les systèmes à usage général. L'environnement d'exécution génératif constitue une infrastructure persistante, sur laquelle la sécurité est superposée en couches. Cette architecture est optimisée pour l'amplitude, la disponibilité et la réactivité.

Mais cette même architecture comporte une limitation structurelle qu'aucun investissement en sécurité en aval ne peut pleinement résoudre : si le modèle génère du contenu halluciné, ce contenu doit être détecté et filtré avant la livraison. La détection est incomplète. Même les systèmes de détection les mieux dotés en ressources et les plus soigneusement conçus - opérant dans des domaines étroits et bien contrôlés avec des données organisées - produisent des scores F1 de détection de l'ordre de 60 à 65 %, ce qui signifie qu'une fraction significative du contenu halluciné n'est pas détectée.

2.1 Le RAG ne résout pas le problème

La génération augmentée par récupération (RAG) est la réponse dominante de l'industrie face aux hallucinations. En connectant un modèle génératif à une base de connaissances externe et en l'obligeant à ancrer ses réponses dans des documents récupérés, le RAG réduit substantiellement les taux d'hallucination par rapport à une génération non ancrée - souvent de 30 à 50 % par rapport à une référence de base.

Cette amélioration est réelle et précieuse. Mais elle est insuffisante pour les environnements institutionnels réglementés. L'étude de l'Université Stanford et de la Yale Law School qui a testé empiriquement les principales plateformes d'IA juridique fondées sur le RAG - Thomson Reuters (Westlaw) et LexisNexis - a constaté des taux d'hallucination de 17 à 33 % dans des tâches réelles de recherche juridique. Ce sont des systèmes d'entreprise bénéficiant d'investissements techniques substantiels, d'une portée de domaine étroite et de corpus juridiques organisés. Ils produisent encore des hallucinations plus d'une fois sur six, et dans le cas de Thomson Reuters, plus d'une fois sur trois.

RAG fails for multiple structural reasons that no tuning or configuration can fully address:

- Échec de récupération - le récupérateur peut manquer des documents critiques ou ne récupérer que des preuves partielles, amenant le modèle à extrapoler à partir de sa mémoire paramétrique.
- Récupération bruitée - les passages récupérés peuvent n'être que vaguement liés, amenant les modèles à synthétiser des affirmations plausibles mais inexactes.
- Mauvaise utilisation du contexte - même en présence de passages corrects, les modèles produisent parfois du contenu non étayé par les références récupérées, voire les contredisant.
- Erreurs de raisonnement - même sous récupération parfaite, les modèles commettent des erreurs logiques : comptage erroné, agrégation incorrecte ou mauvaise interprétation des preuves.

La limitation du RAG en une phrase
Le RAG réduit substantiellement les hallucinations du type 'je ne savais pas', tout en laissant intactes les hallucinations causées par des erreurs de raisonnement, des mauvaises interprétations ou une incapacité à suivre fidèlement les preuves récupérées.

2.2 Les contrôles après génération ne peuvent empêcher ce qui s'est déjà produit

Les garde-fous, les couches de modération, le filtrage des résultats et la révision humaine sont des contrôles appliqués après l'exécution du modèle génératif. Ils peuvent intercepter une fraction des résultats hallucinés avant la livraison. Ils ne peuvent pas effacer le contenu déjà produit et ne peuvent pas tout intercepter.

La formalisation mathématique de la sécurité dans l'art antérieur est révélatrice. Les cadres de sécurité sont définis comme des fonctions appliquées aux résultats du modèle :

$$y' = G(M(y|x))$$

Où M est le modèle, x l'entrée, y le résultat généré et G la fonction de sécurité. Cette formulation exige que M s'exécute - que l'inférence ait déjà eu lieu - pour que G ait quelque chose sur quoi opérer. La sécurité, dans ce paradigme, est par définition post-génération.

Cela inclut le refus. Lorsqu'un système sécurisé refuse de répondre à une question, ce refus est lui-même un résultat de l'inférence générative. Le modèle s'est exécuté. Il a produit un refus. Cela est architecturalement différent d'un système dans lequel le modèle ne s'est jamais exécuté.

La littérature sur le démarrage à froid renforce encore la domination de ce paradigme. Les environnements d'inférence sur toutes les grandes plateformes infonuagiques - AWS SageMaker, Google GKE, Azure ML - sont provisionnés comme une infrastructure persistante ou à mise à l'échelle automatique, dont l'existence est indépendante des décisions d'autorisation individuelles. L'instanciation par requête est traitée comme un antipattern de performance à minimiser, et non comme un mécanisme de sécurité à adopter. La haute disponibilité est une exigence opérationnelle. L'inexistence d'un environnement d'exécution est une condition d'erreur, non un état terminal conçu.

3. L'inférence à exécution conditionnelle : l'alternative architecturale

COMPAiSS introduit un principe architectural catégoriquement différent. Plutôt que de générer d'abord et de filtrer ensuite, COMPAiSS évalue l'autorisation avant qu'une inférence générative puisse se produire.

Cycle de vie d'exécution COMPAiSS
Requête reçue → Le verrou d'exécution évalue les sources institutionnelles → SI autorisé : environnement d'inférence instancié → Réponse générée et transmise → SI non autorisé : aucun environnement d'inférence instancié. Aucun calcul génératif exécuté. Réponse d'échec sécurisé transmise à coût de calcul marginal nul.

Le verrou d'exécution préalable à l'inférence n'est pas un garde-fou, un filtre, une couche de modération ni un contrôle d'accès appliqué à un service d'inférence en cours d'exécution. C'est une condition qui détermine si un environnement d'inférence peut exister pour une requête donnée. En cas d'échec de l'autorisation, aucun calcul génératif n'est effectué. Il n'existe aucune inférence à partir de laquelle des hallucinations pourraient surgir.

Cette distinction n'est pas sémantique. Elle reflète un positionnement fondamentalement différent de l'autorité au sein du cycle de vie d'exécution :

- Dans les systèmes à génération prioritaire : le modèle existe → l'autorisation régit ce qu'il produit.
- Dans les systèmes à exécution conditionnelle : l'autorisation détermine si le modèle existe → le modèle n'existe que lorsque les sources autorisées le confirment.

"COMPAiSS eliminates scope-violation hallucinations by design, and materially reduces - though does not eliminate - generation-quality risks within authorized scope through structured parsing and tightly bounded inference contexts." - COMPAiSS Technical Documentation

Le diagramme ci-dessous illustre la distinction architecturale. Les deux chemins d'exécution commencent par une requête client. Dans les systèmes à génération prioritaire, le modèle est toujours instancié et la sécurité est appliquée ensuite. Dans les systèmes à exécution conditionnelle, le verrou d'autorisation détermine si le modèle existe pour cette requête.

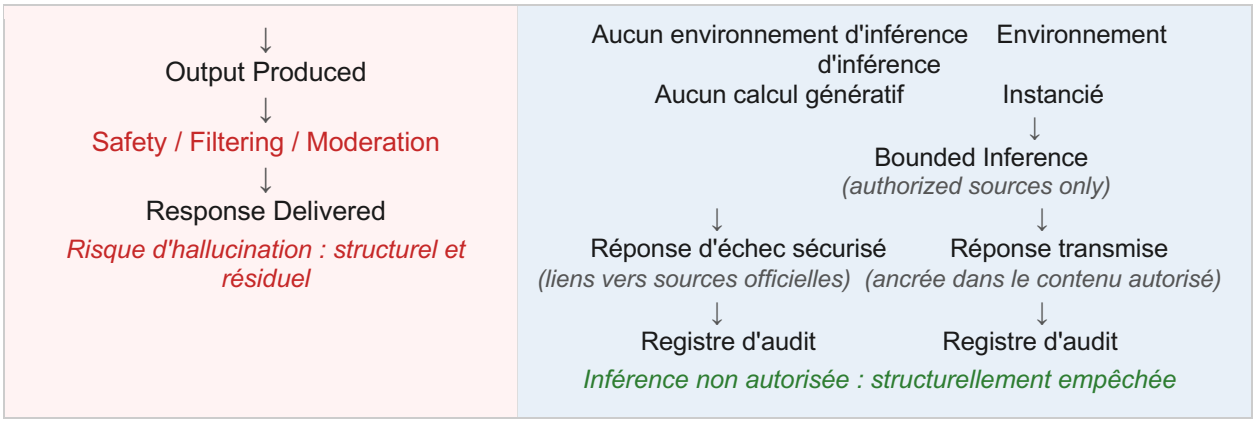

Figure 1 : Comparaison architecturale des systèmes à génération prioritaire (gauche) et de l'inférence à exécution conditionnelle COMPAiSS (droite). Dans les systèmes à génération prioritaire, un modèle génératif s'exécute toujours et la sécurité est appliquée à ses résultats. Dans l'architecture COMPAiSS, le verrou d'autorisation détermine si un environnement d'inférence générative peut exister pour une requête donnée.

3.1 Ce que le verrou évalue

Le verrou d'exécution préalable à l'inférence effectue une découverte automatisée sur la liste de sources approuvées par l'institution - la liste verte (greenlist) - pour déterminer si des documents institutionnels faisant autorité existent pour étayer une réponse à la requête reçue. Trois conditions doivent être satisfaites avant que l'inférence soit autorisée :

- La requête s'inscrit dans la portée définie par l'institution.
- Des documents institutionnels approuvés faisant autorité existent pour étayer une réponse.
- La production d'une réponse serait défendable dans le cadre de gouvernance de l'institution.

Si l'une des conditions n'est pas satisfaite, le système entre dans un état opérationnel non génératif qui est terminal en ce qui concerne l'inférence générative. Aucun poids de modèle n'est chargé. Aucun contexte d'exécution d'inférence n'est alloué. Aucun calcul génératif stochastique n'est effectué. Le client reçoit une réponse structurée d'échec sécurisé - des liens directs vers des sources institutionnelles faisant autorité - à coût de calcul marginal nul.

Si toutes les conditions sont satisfaites, un environnement d'inférence générative est instancié avec un environnement épistémique délimité : le contenu provient exclusivement des pages institutionnelles préautorisées. Le modèle ne raisonne pas sur un corpus vaste et non structuré. Il raisonne uniquement sur les documents approuvés par l'institution.

3.2 Défense en profondeur

COMPAiSS ne prétend pas que le verrouillage d'exécution seul élimine toutes les erreurs de génération possibles. Il affirme quelque chose de plus précis : que le verrouillage d'exécution élimine les hallucinations découlant d'une inférence non autorisée, et réduit matériellement les risques liés à la qualité de génération dans la portée autorisée, grâce à la délimitation étroite du contexte d'inférence.

Within authorized scope, COMPAiSS employs a defense-in-depth architecture:

- Le verrouillage d'autorisation empêche l'exécution non autorisée du modèle principal - éliminant les hallucinations découlant d'une inférence non autorisée sur des requêtes hors portée.
- Les contraintes fondées sur les instructions orientent le comportement du modèle lors de l'inférence autorisée - réduisant la probabilité d'erreurs de génération dans la portée.
- La validation des URL après génération garantit que tous les liens dans les réponses pointent vers des sources institutionnelles autorisées - empêchant la citation de contenu non autorisé.

Il ne s'agit pas d'une séquence de filtres appliqués après un modèle à génération prioritaire. C'est une architecture de gouvernance en couches dans laquelle la contrainte la plus fondamentale - savoir si le modèle s'exécute ou non - est le premier contrôle et le plus important structurellement.

4. Distinction architecturale : art antérieur et analyse du cycle de vie d'exécution

Quatre sondes indépendantes de recherche d'art antérieur - menées à l'aide de Perplexity, Claude et Gemini sur USPTO, WIPO/PCT et Google Patents, complétées par la littérature évaluée par les pairs - ont abouti à un constat cohérent. Dans les références examinées, nous n'avons identifié aucune référence d'art antérieur dans la littérature et le corpus de brevets examinés qui conditionne l'existence d'un environnement d'inférence générative à une autorisation préalable à l'exécution en tant que principe de conception architecturale. Il s'agit d'un constat probatoire dans la portée des sources examinées, et non d'une affirmation d'absence universelle.

L'architecture dominante dans l'ensemble du corpus d'art antérieur considère l'inférence comme un service persistant ou à mise à l'échelle automatique dont l'existence est indépendante des politiques d'autorisation ou de sécurité. L'autorisation détermine si un principal peut invoquer un service déjà en cours d'exécution. La sécurité détermine quels résultats peuvent être livrés. Ni l'une ni l'autre ne détermine si le service existe.

Cette hypothèse n'est pas accidentelle. Elle est structurellement intégrée dans les mathématiques des cadres de sécurité, la conception de l'infrastructure ML infonuagique et les objectifs d'optimisation des systèmes de service LLM. Les références d'art antérieur montrent de multiples formes d'enseignement s'éloignant de l'architecture à exécution conditionnelle :

Type d'éloignement	Description
Mathématique	Les cadres de sécurité définissent la sécurité comme y' = G(M(x)) - une fonction appliquée aux résultats du modèle. Cela exige que M s'exécute, rendant la non-exécution incompatible avec le modèle de sécurité standard.
Architecturale	Les plateformes ML infonuagiques (SageMaker, GKE, Azure ML) provisionnent l'inférence comme une infrastructure persistante ou à mise à l'échelle automatique. L'autorisation régit l'accès ; elle ne contrôle pas l'existence de l'environnement d'exécution.

Type d'éloignement	Description
Opérationnelle	La littérature sur le démarrage à froid traite l'instanciation par requête comme un antipattern de performance. La haute disponibilité est une exigence de niveau de service. La non-existence d'un environnement d'exécution est une condition d'erreur, non un état conçu.
Conceptuelle	Les plateformes infonuagiques découplent intentionnellement la logique d'autorisation du contrôle du cycle de vie de l'infrastructure. Cette séparation est un principe de conception fondamental, non une omission accidentelle.
Systémique	Les cadres de refus - Constitutional AI, Learn-to-Refuse, Bedrock Guardrails - définissent le refus comme un résultat de l'exécution générative, et non comme une non-exécution. La sécurité exige que le modèle s'exécute.

Des évaluations indépendantes par Claude, Gemini et GPT-4 ont chacune confirmé ces constats dans la portée de cet examen. Dans notre examen, aucune référence examinée ne conditionne l'existence de l'inférence à l'autorisation en tant que propriété architecturale conçue. Dans notre examen, aucune référence examinée ne décrit la non-instanciation comme un état terminal conçu du système. La combinaison IAM/RBAC, mise à l'échelle automatique et modération - la combinaison la plus plausible de l'art examiné - ne conduit pas, dans les références d'art antérieur, à une inférence à exécution conditionnelle, et est à plusieurs égards activement découragée par les objectifs d'optimisation déclarés de l'art.

L'architecture COMPAiSS fait l'objet de demandes de brevets en cours d'examen au Canada (OPIC 3 299 174) et aux États-Unis (USPTO 19/455,963). Les revendications indépendantes décrivent un système dans lequel un environnement d'inférence générative n'est instancié que lorsqu'un état d'autorisation existe, et dans lequel, en l'absence d'état d'autorisation, aucun environnement d'inférence générative n'est instancié, aucun contexte d'inférence exécutable n'existe et aucun calcul génératif stochastique n'est effectué.

5. Performance empirique : le déploiement à Service Canada

L'assistant d'information par intelligence artificielle de Service Canada (SCAI) - déployé à compaiss.ca/service-canada/ - est la première mise en oeuvre institutionnelle de l'architecture d'inférence à exécution conditionnelle COMPAiSS dans un contexte de prestation de services gouvernementaux fédéraux. Il offre des capacités de requête en langage naturel pour les programmes de Service Canada, notamment l'assurance-emploi, le Régime de pensions du Canada, la Sécurité de la vieillesse, le Supplément de revenu garanti, le numéro d'assurance sociale et les services connexes.

Les tests empiriques effectués le 26 mai 2026 ont produit les résultats suivants :

5.1 Performance bilingue

Une requête en français sur l'admissibilité à la Sécurité de la vieillesse et les montants des prestations a retourné une réponse complète et exacte en français, ancrée exclusivement dans

des sources du gouvernement du Canada en français. Sept URL /fr/ distinctes ont été citées dans la réponse, couvrant :

- securite-vieillesse (page principale de la SV)
- securite-vieillesse/admissibilite (admissibilité)
- supplement-revenu-garanti (SRG)
- supplement-revenu-garanti/montant-prestation (montants du SRG)
- securite-vieillesse/demande (demande)
- securite-vieillesse/paiements (calendrier des paiements)

Le matériel source cité a été tiré directement des pages du gouvernement du Canada en français - et non traduit de l'anglais. Cela confirme que le système ne traduit pas des réponses en anglais vers le français, mais ancre les requêtes en français dans des sources faisant autorité en français. Cela satisfait directement l'exigence de la Loi sur les langues officielles en matière d'équivalence de la qualité de service.

5.2 Qualité des réponses

Les réponses démontrent une livraison structurée en langage clair, adaptée à la population diversifiée de clients de Service Canada, notamment :

- Des conditions d'admissibilité au programme claires, tirées directement des pages sources faisant autorité
- La citation explicite des URL des sources du gouvernement du Canada dans toutes les réponses
- Les numéros de téléphone ATS (pour les clients sourds et malentendants) inclus dans chaque réponse contenant des informations de contact
- Des prochaines étapes concrètes et une entrée de suivi pour des conseils personnalisés
- Un transfert approprié au personnel de Service Canada pour les requêtes complexes ou sensibles

5.3 Audit d'accessibilité WCAG 2.1 AA Phase 1 - Complété

Un audit d'accessibilité automatisé (axe-core via Playwright Chromium sans interface graphique) a été exécuté directement sur le code source de l'interface en production. Sept violations ont été identifiées lors de deux phases d'audit et toutes ont été corrigées :

- Contraste des couleurs du bouton Soumettre - résolu en modifiant le poids de police de 600 à 700, abaissant le seuil WCAG applicable à 3,0:1. La couleur rouge de la marque Service Canada (#ff0000) est préservée sans modification.
- Repère <main> manquant, H1 manquant et contenu hors des repères - résolu par deux modifications de balises HTML.
- Contraste du texte d'espace réservé - résolu par l'ajout d'une règle CSS ::placeholder (couleur #595959, rapport 6,89:1).
- Région ARIA en direct ajoutée au conteneur de réponse - les lecteurs d'écran annoncent désormais l'état de chargement et l'arrivée de la réponse.
- Les emoji décoratifs dans les résultats sont enveloppés dans des balises span aria-hidden - les lecteurs d'écran n'annoncent plus les descriptions d'emoji avant les titres.
- Texte de divulgation de l'IA ajouté : 'Ceci est un assistant IA. Il fournit des informations uniquement et ne prend pas de décisions concernant vos prestations.'

Quatorze règles d'accessibilité ont été validées lors de l'audit initial, notamment l'étiquetage des formulaires, l'attribution de la langue, la mise à l'échelle du zoom et les vérifications d'accessibilité au clavier. La révision manuelle de la Phase 2 - navigation au clavier en direct, tests de lecteur d'écran NVDA/VoiceOver et vérification des cibles tactiles mobiles - constitue le prochain jalon d'accessibilité prévu.

5.4 Performance multilingue

Au-delà du routage bilingue en langues officielles, l'interface SCAI a été testée avec une requête en langage naturel soumise en tagalog - la langue de plus de 900 000 Canadiens philippins, l'une des communautés d'immigrants à la croissance la plus rapide au Canada et une population importante de clients de Service Canada.

La requête 'Paano ako mag-apply para sa Employment Insurance kung nawalan ako ng trabaho ?' (Comment puis-je demander l'assurance-emploi si j'ai perdu mon emploi ?) a retourné une réponse complète et exacte en tagalog. La réponse comprenait :

- Les taux d'AE 2026 corrects (gains maximaux assurables de 68 900 $, prestation maximale de 729 $/semaine)
- La formule de calcul correcte de 55 % avec des exemples résolus
- Des exemples corrects de taux de chômage régionaux avec les exigences réelles en heures
- Des informations correctes sur la dispense du délai de carence (du 30 mars au 11 avril 2026)
- Le processus de demande étape par étape en tagalog naturel
- Le numéro ATS (1-800-529-3742) pour les clients sourds et malentendants
- Des liens vers les pages sources du gouvernement du Canada (en anglais, car il n'existe pas de pages gouvernementales en tagalog)

La réponse démontre le parcours de traduction multilingue - distinct du parcours de routage bilingue vers les sources utilisé pour les requêtes en français. Pour les langues autres que le français, SCAI traduit la requête aux fins d'évaluation de la gouvernance, récupère à partir de sources autorisées en anglais et livre la réponse dans la langue d'origine du client. Le verrou de gouvernance fonctionne de manière identique quelle que soit la langue de la requête.

Importance de la gouvernance multilingue
Un travailleur canado-philippin qui perd son emploi et cherche des conseils sur l'AE peut recevoir des informations exactes, actuelles et étayées par des sources en tagalog à toute heure, sans attente de 41 minutes. Le système ne peut pas halluciner les taux d'AE qu'il a récupérés à partir de la grille tarifaire officielle de 2026. L'architecture de gouvernance ne fait aucune distinction entre une requête en anglais et une requête en tagalog - l'autorisation est évaluée en anglais par rapport aux sources vérifiées, et la réponse est transmise dans la langue utilisée par le client.

5.5 L'expérience opérationnelle d'un déploiement COMPAiSS

L'architecture décrite dans les sections précédentes se comprend le mieux par le biais de son fonctionnement du point de vue des trois parties qui interagissent avec elle : les clients, les administrateurs institutionnels et le registre de gouvernance.

L'expérience client

Un client visite une page de Service Canada et voit l'interface de requête. Il saisit une question en langage courant - dans n'importe quelle langue. Si sa question s'inscrit dans la portée de la liste verte, il reçoit une réponse structurée en langage clair, ancrée exclusivement dans les pages sources du gouvernement du Canada, avec des liens directs vers ces sources et, le cas échéant, des informations de contact comprenant les numéros ATS pour les clients sourds et malentendants. L'expérience est immédiate, disponible en tout temps et ne nécessite aucune connaissance spécialisée des noms de programmes ou de la terminologie des politiques.

Si leur question est hors portée - parce qu'aucune source autorisée n'y répond ou parce qu'elle concerne un programme absent de la liste verte - ils reçoivent un message clair en langage courant les dirigeant vers le personnel de Service Canada par les canaux existants. Il n'y a pas de réponse partielle ambiguë, pas de contenu halluciné et pas de non-réponse confuse. Le mode d'échec est transparent et immédiatement actionnable.

Gestion de la liste verte : contrôle institutionnel en pratique

La liste verte - la liste de sources approuvées par l'institution qui définit la portée autorisée du système - est gérée via un tableau de bord de liste verte accessible par navigateur aux administrateurs désignés de Service Canada. Par le biais du tableau de bord, les administrateurs peuvent ajouter ou supprimer immédiatement des URL de sources autorisées, consulter l'historique complet des modifications et revenir à toute version précédente. Les modifications prennent effet en quelques secondes. Aucune intervention de développeur n'est requise pour les mises à jour de routine de la portée.

Lorsque les pages de programmes du gouvernement du Canada sont mises à jour - nouvelle politique publiée, taux de prestations révisés, critères d'admissibilité modifiés - les administrateurs de Service Canada signalent la modification par le biais du processus de rétroaction. La liste verte est mise à jour en conséquence. Étant donné que SCAI récupère le contenu de manière dynamique à partir des pages Web actuelles du gouvernement du Canada plutôt que de maintenir un corpus mis en cache, les pages mises à jour se reflètent dans les réponses immédiatement après que la liste verte est confirmée comme étant à jour. Si une page source est supprimée ou si son URL change, le verrou échouera à la récupérer et ne générera pas de réponse à partir de cette source : le client est dirigé vers le personnel de Service Canada plutôt que de recevoir des informations périmées.

Gouvernance et piste d'audit

Chaque interaction génère un registre de gouvernance structuré capturant : la requête reçue, la décision du verrou (autorisée ou refusée), les sources récupérées et analysées, la version du modèle utilisée et si une réponse a été transmise. Ce registre est disponible pour les administrateurs de Service Canada et fournit une chaîne de responsabilité complète et vérifiable pour chaque interaction assistée par l'IA.

Pour les requêtes refusées - environ 40 % dans un déploiement institutionnel typique - le registre documente la raison de l'échec du verrou et la réponse d'échec sécurisé transmise. Pour les requêtes autorisées, le registre retrace la réponse jusqu'aux pages sources spécifiques consultées. À tout moment, un administrateur institutionnel peut vérifier exactement ce que le système a dit et pourquoi. Il ne s'agit pas d'une capacité d'audit probabiliste ou approximative. C'est un registre déterministe de chaque décision prise par le système.

Escalade et supervision humaine

SCAI est un canal supplémentaire, non un remplacement de la prestation de services humains. Chaque réponse dirige les clients vers le personnel de Service Canada pour obtenir une aide supplémentaire. Les améliorations prévues comprennent une option de connexion à un agent à

l'initiative du client et une invitation automatique de transfert humain après plusieurs échanges consécutifs, garantissant que les clients nécessitant un soutien humain y sont activement dirigés. Un client qui reçoit une réponse incorrecte ou insuffisante - ce qui est possible dans la portée autorisée, comme le discute la Section 6 - peut contacter immédiatement le personnel de Service Canada par les canaux existants, où le personnel peut examiner l'interaction à l'aide du registre d'audit.

5.6 Déploiements institutionnels supplémentaires

Au-delà du déploiement fédéral à Service Canada, COMPAiSS est en déploiement actif dans deux universités de recherche canadiennes :

- Université Dalhousie - requêtes relatives aux affaires étudiantes et aux politiques académiques. Dalhousie a mené une évaluation institutionnelle formelle et a sélectionné COMPAiSS à l'issue d'une évaluation concurrentielle. Le déploiement couvre le bien-être étudiant et le soutien en santé mentale, le statut académique, les appels de notes, les procédures d'adaptation et les requêtes connexes de services aux étudiants à enjeux élevés.
- Université McGill - requêtes relatives aux admissions au premier cycle et au statut académique. Le déploiement McGill traite les requêtes complexes sur les délais d'admission, les conditions spécifiques à la faculté en matière de probation académique, les exigences pour les candidats en transfert et les informations sur les notes de coupure pour les candidats des écoles secondaires de l'Ontario et du Québec. Les tests ont confirmé que lorsque les sources autorisées contiennent la réponse, COMPAiSS les cite avec précision ; lorsqu'elles ne la contiennent pas, le système reconnaît explicitement la lacune et dirige l'étudiant vers le bureau approprié - sans halluciner une réponse plausible.

Ces déploiements confirment que l'architecture à exécution conditionnelle se généralise à tous les contextes institutionnels. Le verrou de gouvernance, la gestion de la liste verte, la piste d'audit et le comportement d'échec sécurisé fonctionnent de manière identique, que le déploiement serve des clients de programmes de prestations fédéraux, des étudiants diplômés naviguant dans des procédures d'intégrité académique ou des candidats au premier cycle avec des questions d'admission urgentes.

Le test de comportement aux limites de McGill
Interrogé sur la date limite de soumission des documents justificatifs pour les candidats en transfert provenant d'universités hors Canada pour l'automne 2026, le déploiement McGill a fourni une réponse précise là où la source autorisée en confirmait une - le 30 avril 2026, citée directement depuis la page officielle des notes de coupure pour l'Ontario de McGill, avec la mise en garde de l'institution indiquant qu'il s'agit de points de référence historiques et non de minimums garantis. Interrogé sur une question connexe pour laquelle la source autorisée ne précisait pas de date limite, le système l'a reconnu explicitement et a dirigé l'étudiant vers le portail du candidat et le bureau des admissions. Aucune date n'a été fabriquée. La frontière entre ce qui est autorisé et ce qui ne l'est pas est appliquée structurellement, non probabilistement.

6. Le fondement pour les environnements institutionnels réglementés

L'argument standard de l'industrie en faveur de la gestion du risque d'hallucination - selon lequel le RAG plus les garde-fous plus la supervision humaine suffisent - repose sur une hypothèse de tolérance à l'erreur : qu'un certain taux résiduel de résultats incorrects est acceptable compte tenu des avantages du déploiement.

Cette hypothèse est défendable pour l'IA grand public à usage général. Elle n'est pas défendable pour les environnements institutionnels réglementés où les résultats des systèmes d'IA affectent l'accès des citoyens aux prestations, aux droits légaux, aux droits aux soins de santé ou au statut éducatif.

Les orientations réglementaires du gouvernement du Québec illustrent les enjeux : les citoyens qui se fient aux informations générées par l'IA pour prendre des décisions financières ou relatives aux soins de santé peuvent faire face à de graves conséquences lorsque ces informations sont incorrectes. Les tribunaux français ont émis des avertissements judiciaires formels à la fin de 2025 concernant le contenu juridique halluciné par l'IA parvenant aux procédures judiciaires. Le GAO des États-Unis a documenté que les cas d'utilisation de l'IA par les agences ont été multipliés par neuf de 2023 à 2024, avec plus de 85 % des cas d'utilisation déployés à fort impact dépourvus de la documentation requise sur l'atténuation des risques.

Cet argument ne rejette pas le RAG ou les architectures à génération prioritaire. Pour l'IA à usage général - recherche de connaissances en entreprise, génération de contenu, outils de recherche généralistes, soutien à la clientèle dans des domaines ouverts - ces architectures servent bien leurs objectifs. Les ingénieurs qui ont construit des systèmes RAG soigneux et bien gouvernés ont accompli un travail véritablement utile. La question n'est pas de savoir si ces systèmes sont bons pour ce pour quoi ils sont conçus. La question est de savoir s'ils répondent aux obligations de gouvernance spécifiques qu'imposent les contextes institutionnels réglementés : des obligations de fournir des informations faisant autorité, traçables et délimitées aux citoyens dont les droits, la santé ou le statut juridique peuvent dépendre de ce qu'on leur dit.

Dans ce contexte, l'approche de gestion des risques face aux hallucinations crée une exposition institutionnelle et juridique que les institutions elles-mêmes ne peuvent facilement assumer. Un organisme gouvernemental qui déploie un système d'IA fondé sur le RAG et documente un taux résiduel d'hallucination de 17 à 33 % - le taux empiriquement confirmé pour les principaux systèmes d'IA juridique d'entreprise - ne peut pas représenter ce système comme fournissant des informations exactes et faisant autorité au public.

L'inférence à exécution conditionnelle modifie ce calcul. En éliminant la condition structurelle dans laquelle surviennent les hallucinations découlant d'une inférence non autorisée, COMPAiSS permet aux institutions de représenter leurs systèmes d'IA comme délimités, vérifiables et gouvernables d'une manière que les systèmes à génération prioritaire ne peuvent égaler :

- Chaque réponse est traçable jusqu'aux sources institutionnelles autorisées spécifiques.
- Les requêtes non autorisées ne produisent aucun contenu généré par l'IA - uniquement des liens vers des sources officielles.
- L'architecture de gouvernance est déterministe et documentable : l'institution contrôle quelles sources font autorité, définit la limite de portée, et le système applique les deux structurellement.
- La piste d'audit capture chaque requête, chaque décision du verrou et chaque source consultée.

- Aucune donnée personnelle n'est collectée, traitée ou conservée.

6.1 Traitement des erreurs de génération dans la portée autorisée

Un évaluateur averti observera à juste titre que le verrouillage d'exécution dans COMPAiSS empêche les hallucinations hors portée, mais n'élimine pas toutes les erreurs de génération dans la portée autorisée. Cette observation est exacte. Toutefois, la nature des erreurs dans la portée est catégoriquement différente du risque d'hallucination dans les systèmes sans verrou - et la comparaison pertinente n'est pas entre COMPAiSS et un système d'information parfait.

Dans la portée autorisée, chaque réponse COMPAiSS est tirée exclusivement des pages sources institutionnelles vérifiées du gouvernement du Canada. Le verrou d'exécution préalable à l'inférence empêche la génération de toute réponse à moins qu'au moins une source vérifiée ne soit disponible. Par conséquent, le système ne peut pas fabriquer des sources non autorisées ni inventer des détails de programme. Les erreurs dans la portée sont donc principalement navigationnelles ou interprétatives, et non fabricatoires ou décisionnelles. Dans le pire des cas, un client peut être dirigé vers une page autorisée du gouvernement du Canada moins pertinente plutôt que vers la plus directement applicable. De telles erreurs sont de faible impact et facilement corrigeables : les clients peuvent affiner leur requête, consulter d'autres sources liées ou contacter le personnel de Service Canada par les canaux existants. Il est important de noter que SCAI n'a aucune autorité décisionnelle : il ne détermine pas l'admissibilité, ne calcule pas les montants des prestations et ne statue pas sur les demandes, et ses résultats ne créent ni ne modifient les droits des clients.

Dans le pire des cas, une réponse incorrecte de SCAI est navigationnelle ou interprétative, et non fabricatoire ou décisionnelle : un client peut être dirigé vers une page autorisée du gouvernement du Canada moins pertinente. Dans le pire des cas, une réponse d'IA sans verrou est une réponse confiante et spécifique mais erronée - un montant en dollars, une règle d'admissibilité ou un délai halluciné - sur laquelle un client vulnérable agit sans savoir que l'information est fausse.

Ce profil du pire des cas doit être évalué par rapport à l'alternative réaliste - ce qui se produit régulièrement lorsque les clients ne peuvent pas obtenir de réponses immédiates de Service Canada et se tournent vers des systèmes d'IA à usage général tels que ChatGPT, Gemini ou Copilot. Ces systèmes ne sont pas limités aux sources vérifiées du gouvernement du Canada. Ils sont empiriquement et architecturalement documentés pour halluciner des détails spécifiques aux programmes, notamment les règles d'admissibilité, les montants des prestations et les délais de demande - produisant des réponses confiantes et spécifiques mais incorrectes que des clients vulnérables peuvent utiliser sans savoir que l'information est fausse. Ce n'est pas un pire cas théorique. C'est une conséquence routinière de l'inférence sans verrou appliquée à des questions institutionnelles.

La comparaison n'est donc pas entre COMPAiSS et la perfection. Elle est entre deux profils d'échec distincts :

- Un système dont le pire cas est un lien autorisé moins pertinent - immédiatement ou facilement corrigeable, sans autorité décisionnelle et n'introduisant aucun nouveau risque décisionnel ; et

- Un système dont le pire cas est un montant en dollars, une règle d'admissibilité ou un délai halluciné qu'un citoyen utilise pour prendre une décision financière ou administrative importante.

Une troisième considération renforce cette asymétrie. En résolvant avec précision et immédiatement les requêtes d'information routinières, SCAI libère les agents et conseillers de Service Canada pour se concentrer sur les cas complexes, les situations sensibles et les appels - les interactions où le jugement humain n'est pas seulement préférable mais essentiel. Cette réallocation augmente la probabilité de résultats corrects pour l'ensemble de la clientèle de Service Canada, et non seulement pour les clients qui utilisent SCAI. Le bénéfice de gouvernance de l'inférence à exécution conditionnelle s'étend ainsi au-delà des interactions qu'elle traite directement aux interactions qu'elle permet au personnel humain de gérer plus efficacement.

7. Compromis, limitations et choix de conception

L'inférence à exécution conditionnelle est une architecture optimisée pour un ensemble spécifique de contraintes. Comme tout choix architectural, elle implique des compromis. Comprendre ce qu'elle sacrifie - et pourquoi ces sacrifices sont acceptables dans le contexte de déploiement cible - est essentiel pour évaluer si c'est le bon choix pour une institution donnée.

7.1 Les décisions sur les limites de portée requièrent un jugement institutionnel continu

La liste verte est le mécanisme par lequel une institution définit ce à quoi le système peut répondre. C'est une force de gouvernance - l'institution conserve un contrôle complet sur la portée - mais c'est aussi une responsabilité opérationnelle continue. Les listes vertes doivent être activement maintenues à mesure que les programmes évoluent, que les politiques changent et que de nouveaux services sont ajoutés. Une requête qui tombe en dehors de la portée actuelle de la liste verte ne recevra pas de réponse générée par l'IA, même si la réponse est disponible sur une page du gouvernement du Canada qui n'a pas encore été ajoutée.

Les institutions qui déploient COMPAiSS doivent désigner des administrateurs responsables de la maintenance de la liste verte et établir des processus pour signaler les lacunes de portée au fur et à mesure qu'elles apparaissent. Le tableau de bord de liste verte rend cela opérationnellement simple, mais le jugement institutionnel sur ce qui appartient à la portée - et ce qui n'y appartient pas - reste une responsabilité humaine qui ne peut pas être déléguée au système lui-même.

7.2 La qualité de génération dans la portée n'est pas garantie

Le verrouillage d'exécution élimine les hallucinations de violation de portée. Il n'élimine pas les risques liés à la qualité de génération dans la portée autorisée. Lorsque le verrou autorise une requête et que le système génère une réponse à partir du contenu des sources autorisées, le modèle génératif sous-jacent peut encore mal interpréter des preuves, agréger incorrectement des informations provenant de plusieurs sources ou produire des réponses techniquement ancrées mais pratiquement trompeuses. C'est pourquoi COMPAiSS emploie une architecture de défense en profondeur dans la portée autorisée : les contraintes fondées sur les instructions,

l'analyse structurée des sources et la validation des URL après génération réduisent toutes - sans les éliminer - le risque de génération dans la portée.

Les institutions qui déploient COMPAiSS devraient maintenir un parcours d'escalade humaine, surveiller la qualité des réponses via le registre d'audit et mettre à jour les configurations de la liste verte et l'architecture des instructions lorsque des problèmes de qualité systématiques sont identifiés. Le système est conçu pour soutenir cette boucle de gouvernance, non pour la remplacer.

7.3 La portée délimitée est une caractéristique, non une limitation

L'objection la plus courante à l'inférence à exécution conditionnelle est qu'elle ne peut pas répondre aux questions en dehors de sa portée autorisée. C'est exact. C'est aussi l'objectif. COMPAiSS n'est pas conçu pour répondre à n'importe quelle question. Il est conçu pour répondre aux questions qu'une institution l'a autorisé à répondre, à partir de sources approuvées par l'institution, de manières que l'institution peut vérifier et défendre.

Pour les institutions dont les obligations de gouvernance de l'IA nécessitent exactement ce comportement délimité et vérifiable - et où les conséquences des réponses non autorisées sont supportées par les citoyens - ce n'est pas une limitation. C'est la spécification de conception. Les 40 % de requêtes pour lesquelles l'inférence est entièrement refusée ne représentent pas des défaillances du système. Elles représentent le verrou de gouvernance fonctionnant comme prévu : diriger les clients vers des canaux humains pour les questions qui tombent en dehors de la portée d'information approuvée par l'institution.

7.4 Considérations relatives à la latence

L'étape d'autorisation préalable à l'inférence ajoute une petite latence par rapport à l'inférence sans protection. Dans le déploiement de Service Canada, cela inclut un passage de notation de la liste verte par rapport aux URL des sources autorisées et une requête au moteur de recherche personnalisé Google pour identifier les pages sources pertinentes. En pratique, le temps de réponse total pour les requêtes autorisées est comparable à d'autres déploiements d'assistants IA, car les étapes d'autorisation et de récupération se produisent en parallèle avec le temps de traitement perçu par l'utilisateur. Les requêtes refusées retournent plus rapidement que les requêtes autorisées, car aucune inférence n'est exécutée.

Pour les applications nécessitant des temps de réponse inférieurs à la seconde à très grand volume, l'architecture reste applicable, mais la configuration du déploiement - taille de la liste verte, portée de la récupération, stratégie de mise en cache - doit être optimisée en conséquence. C'est une question de configuration, non une contrainte architecturale.

7.5 Le compromis fondamental

Pourquoi le compromis en vaut la peine pour les contextes institutionnels réglementés
L'inférence à exécution conditionnelle sacrifie l'amplitude du domaine ouvert en échange d'un comportement délimité, vérifiable et gouvernable. Pour l'IA à usage général, c'est le mauvais compromis. Pour les environnements institutionnels réglementés où les citoyens s'appuient sur des informations assistées par l'IA pour prendre des décisions concernant leurs prestations, leur santé, leur statut juridique ou leurs droits à l'éducation, c'est le seul compromis qui satisfait aux obligations de gouvernance de l'institution.

8. Comparaison avec les approches existantes

Critère	LLM à usage général	Système RAG	Garde-fous / Filtrage	COMPAiSS (à exécution conditionnelle)
Prévention des hallucinations	Aucune - génère librement	Réduit de 30-50 %	Détecte/filtre après génération	**Élimine les hallucinations découlant d'une inférence non autorisée**
Taux résiduel d'hallucination	Élevé	17-33 % (entreprise)	Dépend de la qualité du détecteur	**Zéro hors portée**
Vérifiabilité de la gouvernance	Faible	Modérée	Modérée	**Élevée - chaque décision consignée**
Coût d'infrastructure	Faible à modéré	90 000-200 000 $/an (100 000 req.)[1]	S'ajoute au coût RAG	**15 000-30 000 $/an (100 000 req.)[1]**
Contrôle institutionnel	Aucun	Partiel	Partiel	**Complet - liste verte appartenant à l'institution**
Langues officielles	Traduction seulement	Variable	Variable	**Routage bilingue vers les sources, vérifié**
Confidentialité - données personnelles	Souvent conservées	Souvent conservées	Dépend du système	**Sans état - rien n'est conservé**
Statut du brevet	S. O.	S. O.	S. O.	**OPIC 3 299 174 / USPTO 19/455,963**

[1] Les estimations supposent environ 100 000 requêtes assistées par l'IA annuellement. Les chiffres RAG reflètent les coûts d'exploitation annuels totaux, incluant le calcul d'inférence, l'infrastructure de base de données vectorielle et les outils de gouvernance compensatoires. Les chiffres COMPAiSS reflètent les coûts d'API d'inférence, la récupération de la liste verte et l'hébergement, avec environ 40 % des requêtes refusées à coût marginal nul. Les coûts réels varient selon le volume, le modèle et la configuration. Ces chiffres sont indicatifs et directionnels, non contractuels.

9. Conclusion

L'aperçu le plus important de cet article n'est pas que les taux d'hallucination peuvent être réduits. C'est que l'existence d'un environnement d'inférence générative peut elle-même devenir contrôlée par la gouvernance. Ce changement - de la gestion de ce qu'un modèle produit à la gouvernance du fonctionnement d'un modèle - est la distinction fondamentale entre l'atténuation des hallucinations et la prévention des hallucinations par conception architecturale.

La persistance des hallucinations de l'IA malgré toutes les approches d'atténuation actuelles n'est pas un échec de l'effort d'ingénierie. C'est une conséquence d'une hypothèse architecturale commune : que l'inférence générative s'exécute d'abord et que la sécurité est appliquée ensuite. Chaque système RAG, chaque garde-fou, chaque couche de modération, chaque flux de révision humaine opère sur des résultats que le modèle génératif a déjà produits. Les hallucinations découlant d'une exécution d'inférence non autorisée sont structurellement inévitables dans ce paradigme. La détection et le filtrage en interceptent une fraction.

L'inférence à exécution conditionnelle traite ce problème au niveau architectural. En conditionnant l'existence d'un environnement d'inférence générative à la confirmation préalable de sources institutionnelles autorisées, COMPAiSS empêche la condition structurelle dans laquelle surviennent les hallucinations découlant d'une exécution d'inférence non autorisée. Dans la portée autorisée, les risques liés à la qualité de génération subsistent et sont gérés par des contrôles de défense en profondeur. C'est l'affirmation honnête et précise.

Cette approche n'a aucune référence d'art antérieur identifiée dans la littérature et le corpus de brevets examinés qui divulgue une architecture de cycle de vie d'exécution équivalente. Elle est en cours d'examen de brevet au Canada et aux États-Unis. Elle a été validée empiriquement dans trois déploiements institutionnels canadiens - un contexte de prestation de services gouvernementaux fédéraux servant les Canadiens dans les deux langues officielles, et deux universités de recherche opérant dans des domaines à enjeux élevés de services aux étudiants et d'admissions - et inscrite sur la Liste de sources en intelligence artificielle du gouvernement du Canada (AchatsCanada).

Pour les environnements institutionnels réglementés - gouvernements, hôpitaux, services juridiques, universités - où les conséquences de la désinformation générée par l'IA sont supportées par les citoyens plutôt que par les systèmes, la question n'est pas de savoir si l'inférence à exécution conditionnelle est préférable. La question est de savoir si les obligations de gouvernance de ces institutions peuvent être satisfaites sans elle.

Références

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & 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

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

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

Huang, L., et al. (2024). A Survey on Hallucination in Large Language Models. ACM Computing Surveys.

Shuster, K., et al. (2023). Retrieval Augmentation Reduces Hallucination in Conversation. arXiv:2104.07567.

Harvey, F. (2026). COMPAiSS: Execution-Gated Inference Architecture. Patent applications: CIPO 3,299,174 (Canada); USPTO 19/455,963 (United States).

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

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

Prevail AI. (2024). Stanford Study Reveals Challenges of RAG in Legal AI Tools. blog.prevail.ai.

Treasury Board of Canada Secretariat. (2019, amended 2023). Directive on Automated Decision-Making. Canada.ca.