Autoriser les données probantes avant l'inférence

Positionnement de COMPAiSS dans le paysage émergent des architectures d'IA axées sur la gouvernance

ÉBAUCHE — Document de positionnement et de différenciation

[Nom de l'auteur] · COMPAiSS · Juillet 2026 · compaiss.ca

Résumé

Les établissements font face à un problème structurel croissant : le public pose à des systèmes d'intelligence artificielle des questions à leur sujet, et rien ne limite ce que ces systèmes utilisent comme données probantes pour y répondre. Une famille grandissante d'architectures y répond en insérant la gouvernance dans le pipeline d'exécution de l'IA plutôt qu'en s'en remettant au seul comportement du modèle. Le présent document positionne COMPAiSS — un assistant institutionnel déployé dans lequel l'établissement autorise, avant que l'inférence ne commence, les sources probantes sur lesquelles une réponse peut se fonder — au sein de cette famille. Nous organisons le paysage selon deux axes : le moment où le contrôle s'applique (à l'entraînement, avant l'inférence, pendant la génération ou après la génération) et l'objet de l'autorisation (valeurs, catégories de contenu, flux de dialogue, actions et capacités, ou données probantes). Par rapport aux voisins architecturaux les plus proches — méthodes d'entraînement constitutionnelles, classificateurs de sauvegarde en entrée-sortie, garde-fous programmables à l'exécution et conceptions fondées sur les capacités comme CaMeL — COMPAiSS occupe une position distincte définie par une combinaison précise de propriétés : à notre connaissance, il s'agit de l'implémentation déployée dans laquelle l'objet de l'autorisation préalable à l'inférence est la donnée probante institutionnelle, où la partie autorisante est l'établissement lui-même, où l'application est déterministe et en mode fermé par défaut, et où les réponses conservent une traçabilité énoncé par énoncé. Nous ne prétendons pas être le seul effort déplaçant la gouvernance en amont ; une terminologie convergente est apparue indépendamment dans de récents rapports techniques auto-publiés. Nous caractérisons plutôt COMPAiSS comme une implémentation d'une approche architecturale fondée sur l'autorisation préalable, qui se distingue par l'autorisation des données probantes par l'établissement avant l'inférence, une barrière d'exécution déterministe avec refus gouverné en mode fermé par défaut, la traçabilité de chaque énoncé et un ensemble d'évaluations comparatives en déploiement.

1. Introduction et objectif

Les assistants d'IA à usage général répondent désormais à des questions propres aux établissements — consignes de préparation hospitalière, règles d'admissibilité aux prestations, politiques universitaires — pour des millions de personnes qui ne visitent jamais le site Web de l'établissement concerné. Les réponses sont souvent fluides, souvent utiles et, dans une proportion non négligeable, erronées d'une manière que l'établissement ne voit jamais et ne peut corriger. Le risque institutionnel n'a rien d'abstrait : un organisme d'appel inventé, une mauvaise entrée d'hôpital ou une règle d'admissibilité périmée entraînent des conséquences opérationnelles et juridiques qui retombent sur l'établissement décrit, et non sur le système qui le décrit.

Un corpus de travaux en développement rapide répond à cette classe de problèmes en ajoutant des couches de contrôle aux systèmes d'IA. Ces efforts diffèrent considérablement quant à l'endroit où le contrôle s'applique et à ce qui, précisément, est contrôlé. L'objectif du présent document est de positionner avec précision un système déployé — COMPAiSS — dans ce paysage : énoncer sa proposition architecturale en termes vérifiables, nommer honnêtement ses voisins les plus proches et préciser ce qui le distingue sans surestimer sa nouveauté.

La thèse de positionnement de ce document est volontairement modeste dans sa forme et précise dans son contenu : ***COMPAiSS est une implémentation d'une approche architecturale fondée sur l'autorisation préalable, qui se distingue par l'accent mis sur les données probantes autorisées par l'établissement avant l'inférence, une barrière d'exécution déterministe et une méthodologie de préservation des données probantes.***

2. La proposition architecturale, énoncée avec précision

COMPAiSS repose sur un principe organisateur unique : l'établissement détermine ce que l'IA est autorisée à utiliser comme données probantes avant que l'inférence ne commence, et l'inférence ne se produit qu'à l'intérieur de cette frontière probante autorisée. Concrètement, l'architecture comporte cinq propriétés porteuses :

- L'autorisation des données probantes précède l'inférence. Une barrière d'exécution (execution gate) vérifie, avant toute génération, s'il existe des données probantes institutionnelles autorisées pour la question posée. Cette barrière est une propriété structurelle du pipeline, et non une instruction adressée au modèle.
- L'établissement est la partie autorisante. Les sources autorisées sont définies par une liste verte (greenlist) que l'établissement contrôle. L'autorité de gouvernance demeure entre les mains de l'organisation où le système est déployé, et non du développeur du modèle ni du fournisseur de la plateforme.
- L'inférence est confinée aux données probantes autorisées. Le modèle synthétise et raisonne — il ne s'agit pas d'un simple index de recherche —, mais la matière sur laquelle il peut raisonner est bornée ex ante. Les connaissances issues de l'entraînement ne peuvent se substituer aux données probantes institutionnelles manquantes.
- La défaillance est en mode fermé et gouvernée. Lorsque les données probantes autorisées sont insuffisantes, le système déclare la frontière et réoriente vers un canal humain plutôt que de générer à partir de connaissances générales. Le refus est un résultat conçu et légitime.
- Les réponses sont traçables énoncé par énoncé. Chaque affirmation institutionnelle d'une réponse cite la source autorisée dont elle découle, rendant les erreurs identifiables et corrigibles à la source.

Cette proposition est plus étroite qu'un appel général à une conception « gouvernance d'abord », et elle est vérifiable : pour toute réponse donnée, on peut se demander si chaque affirmation institutionnelle se rattache à une source figurant sur la liste verte de l'établissement, et si le système s'abstient

lorsqu'aucune source de ce type n'existe. Des évaluations comparatives appliquant exactement ce test sont résumées à la section 6.

3. Une taxonomie des points de contrôle

Les architectures d'IA axées sur la gouvernance se distinguent le mieux selon deux axes. Le premier est le moment : quand, par rapport à l'inférence, le contrôle s'applique-t-il ? Quatre positions reviennent dans la littérature et dans les systèmes déployés : (i) à l'entraînement, où le contrôle est intégré aux poids du modèle ; (ii) avant l'inférence, où une condition doit être satisfaite pour que la génération ait même lieu ; (iii) pendant la génération, où des garde-fous encadrent le dialogue, les sujets ou les appels d'outils en temps réel ; et (iv) après la génération, où les sorties sont classifiées, filtrées ou corrigées.

Le second axe — moins souvent explicité et, à notre avis, le plus lourd de conséquences — est l'objet de l'autorisation : quelle est la chose qui est permise ou refusée ? Les objets possibles comprennent les valeurs et dispositions du modèle ; les catégories de contenu (taxonomies de sûreté) ; les flux de dialogue et les sujets ; les outils, actions, capacités et flux de données ; et les données probantes — les connaissances sur lesquelles une réponse est autorisée à se fonder. Deux systèmes peuvent tous deux revendiquer une « autorisation avant exécution » tout en autorisant des choses entièrement différentes, avec des conséquences institutionnelles entièrement différentes. L'analyse qui suit applique ces deux axes aux voisins les plus proches recensés dans une revue vérifiée de l'art antérieur.

4. Les voisins architecturaux

4.1 Alignement des valeurs à l'entraînement : Constitutional AI

Constitutional AI (Bai et coll., 2022) entraîne les modèles à partir d'un ensemble explicite de principes, en utilisant la rétroaction de l'IA pour façonner l'innocuité sans étiquetage humain massif. C'est une gouvernance logée dans les poids : la constitution influence la disposition du modèle à chaque inférence ultérieure. Ses forces sont la généralité et l'échelle ; sa limite structurelle, à des fins institutionnelles, est qu'une disposition n'est pas une application contraignante. Un modèle entraîné constitutionnellement répond toujours aux questions propres à un établissement à partir de ce que contient sa distribution d'entraînement, et aucun établissement extérieur au développeur du modèle ne tient la plume. Constitutional AI et l'autorisation des données probantes sont complémentaires plutôt que concurrentes : COMPAiSS peut fonctionner au-dessus de modèles entraînés constitutionnellement tout en ajoutant la frontière probante institutionnelle qui leur fait défaut.

4.2 Classification aux frontières : Llama Guard

Llama Guard (Inan et coll., 2023) est une sauvegarde d'entrée-sortie fondée sur un LLM : un classificateur appliqué à la frontière conversationnelle pour détecter les contenus qui contreviennent à une taxonomie de sûreté. Il illustre la famille du filtrage a posteriori et aux frontières — la gouvernance comme inspection de ce qui entre et de ce qui sort. Deux propriétés le distinguent de l'autorisation des données probantes. Premièrement, son objet est la catégorie de contenu (violence, automutilation, etc.), et non la provenance probante : une réponse fluide et conforme aux politiques qui contient un fait institutionnel inventé passe

sans encombre. Deuxièmement, il est lui-même un modèle, donc probabiliste : ses jugements sont sujets au même comportement distributionnel que celui qu'il est chargé de surveiller.

4.3 Garde-fous programmables à l'exécution : NeMo Guardrails et la gouvernance d'exécution

NeMo Guardrails de NVIDIA (2023 à aujourd'hui) et ses extensions récentes de gouvernance à l'exécution offrent des garde-fous programmables autour des applications LLM : contrôle des sujets, vérifications en entrée-sortie, contraintes de récupération et contrôle de l'invocation d'outils, exprimés dans une configuration au niveau de l'application. Cette famille déplace la gouvernance sensiblement en amont du simple filtrage des sorties et donne aux développeurs de réelles surfaces de contrôle — y compris des rails de récupération qui restreignent les documents entrant dans le contexte. C'est, dans l'esprit, la famille d'outils largement déployée la plus proche de COMPAiSS. Les différences tiennent au siège de l'autorité et aux paramètres par défaut : les garde-fous sont une trousse qu'un développeur peut configurer (y compris de façon permissive), leur application mêle règles déterministes et vérifications par modèle, et le cadre reste agnostique quant à qui détient l'autorité. COMPAiSS est une architecture institutionnelle conçue à dessein, dans laquelle l'autorisation des données probantes est le défaut non négociable, l'établissement est l'autorité définie et le refus en l'absence de données probantes est un résultat conçu plutôt qu'un choix de configuration.

4.4 Contrôle des flux fondé sur les capacités : CaMeL — le voisin architectural le plus proche

CaMeL (Debenedetti et coll., 2025) est, selon notre analyse et selon la revue de l'art antérieur, l'architecture existante la plus proche. En réponse aux attaques par injection d'invites contre les agents LLM, CaMeL sépare le flux de contrôle du flux de données et applique une autorisation fondée sur les capacités : ce qu'un agent peut faire, et quelles données peuvent circuler où, est déterminé par une politique de sécurité explicite avant l'exécution, et non par le jugement du modèle. Il partage avec COMPAiSS les deux engagements les plus profonds — l'autorisation avant l'exécution et une application déterministe située à l'extérieur du modèle.

La distinction réside dans l'objet de l'autorisation. CaMeL autorise des actions et des flux de données ; son modèle de menace est l'instruction injectée qui pousse un agent à poser un geste non autorisé. COMPAiSS autorise des données probantes ; son modèle de menace est la réponse fluide fondée sur quelque chose que l'établissement n'a jamais dit. Les deux retirent au modèle une décision critique, mais ce ne sont pas les mêmes décisions. Un hôpital qui déploie un agent a besoin d'un contrôle des capacités à la CaMeL pour que l'agent ne puisse pas mal agir ; le même hôpital qui déploie un service public de réponses a besoin d'un contrôle des données probantes pour que la réponse ne puisse pas mal affirmer. Les deux approches sont architecturalement compatibles, et une pile institutionnelle mature en contiendra vraisemblablement les deux.

4.5 Terminologie convergente : les rapports sur la gouvernance pré-inférence

Une série de rapports techniques auto-publiés (Akarkach, 2026a, 2026b, 2026c ; Zenodo, non révisés par les pairs) définit la « gouvernance pré-inférence » comme un paradigme dans lequel l'inférence est une

capacité conditionnelle exigeant une autorisation préalable, la non-exécution en mode fermé constituant un état légitime. Nous citons ces rapports pour une raison précise : ils documentent que le vocabulaire de l'autorisation avant l'inférence est apparu indépendamment en plusieurs endroits, ce qui appuie l'idée que la gouvernance remonte en amont comme orientation d'ensemble du domaine. Deux limites encadrent cette citation. Premièrement, ces rapports sont conceptuels plutôt qu'opérationnels : selon leur propre description, ils définissent un principe et n'offrent ni implémentation, ni architecture réalisée, ni évaluation. Deuxièmement, COMPAiSS a été conçu et déployé indépendamment de ces travaux et est décrit tout au long du présent document dans son propre vocabulaire établi — autorisation des données probantes, barrière d'exécution, refus gouverné. La comparaison n'est incluse que pour documenter une convergence de vocabulaire ; la substance architecturale sur laquelle repose COMPAiSS — un mécanisme concret et déployé dont l'objet autorisé est la donnée probante institutionnelle — est établie par les propriétés et les évaluations des sections 2 et 6, et non par cette convergence.

5. Matrice comparative

La matrice ci-dessous applique les deux axes de la section 3 aux voisins vérifiés. Les cellules décrivent chaque approche telle qu'elle est documentée dans ses sources primaires ; les rapports sur la gouvernance pré-inférence y figurent comme principe énoncé plutôt que comme système implémenté, conformément à leur propre description.

Approche	Point de contrôle (moment)	Objet de l'autorisation	Nature de l'application	Principal mode de défaillance visé	Siège de l'autorité
Constitutional AI (Bai et coll., 2022)	À l'entraînement	Valeurs et dispositions comportementales du modèle	Probabiliste (apprise)	Contenus nuisibles ou contraires à l'éthique	Développeur du modèle
Llama Guard (Inan et coll., 2023)	Autour de l'inférence (entrées/sorties)	Catégories de contenu (taxonomie de sûreté)	Probabiliste (classificateur)	Contenus non sûrs franchissant la frontière	Développeur du modèle ou de la plateforme
NeMo Guardrails / Relay (NVIDIA, 2023–2026)	À l'exécution, autour de la génération et des appels d'outils	Flux de dialogue, sujets, invocation d'outils	Garde-fous programmables ; mélange de règles déterministes et de vérifications par modèle	Comportements hors sujet, non sûrs ou non autorisés des outils	Développeur de l'application
CaMeL (Debenedetti et coll., 2025)	Avant l'exécution (flux de contrôle et de données)	Actions, capacités et flux de données d'un agent	Déterministe (fondée sur les capacités)	Injection d'invites ; actions non autorisées	Concepteur du système (politique de sécurité)
Rapports sur la gouvernance pré-inférence (Akarkach, 2026)	Avant l'inférence (principe énoncé)	Légitimité du calcul lui-même (abstrait)	Définitionnelle ; explicitement non opérationnelle	L'inférence non autorisée en tant que telle	Autorité responsable externe (abstrait)
COMPAiSS	Avant l'inférence	Les données probantes — les sources institutionnelles autorisées à fonder une réponse	Barrière d'exécution déterministe ; refus gouverné en mode fermé par défaut	Attribution institutionnelle erronée et faits institutionnels hallucinés	L'établissement déployeur (liste verte)

Tableau 1. Positionnement selon le point de contrôle, l'objet de l'autorisation, la nature de l'application, le mode de défaillance et le siège de l'autorité. Sources : documents primaires énumérés dans les références ; caractérisations vérifiées contre les registres publics en juillet 2026.

6. Ce qui distingue COMPAiSS

Dans ce paysage, COMPAiSS se distingue par la conjonction de cinq propriétés. Aucune n'est individuellement sans précédent sur le plan conceptuel ; à notre connaissance, leur combinaison dans un système déployé et évalué l'est.

- La donnée probante comme objet de l'autorisation. Là où le voisin déterministe le plus proche (CaMeL) autorise des actions et des flux de données, COMPAiSS autorise le fondement probant des réponses — la propriété dont les établissements sont réellement responsables lorsque leur nom figure sur la réponse.
- L'établissement comme partie autorisante. La liste verte place l'autorité de gouvernance entre les mains de l'organisation où le système est déployé. C'est une distinction de siège de contrôle par rapport aux approches au niveau du modèle (autorité du développeur) comme aux trousses de garde-fous (discrétion du développeur d'application).
- Une application déterministe, en mode fermé par défaut. La barrière d'exécution est une structure du pipeline, non un contenu d'invite ni le jugement d'un classificateur. En l'absence de données probantes autorisées, le refus gouverné avec réorientation vers un humain est le résultat conçu — en essais de déploiement, le système déclare explicitement « absent des sources fournies » plutôt que d'importer du matériel externe.
- La préservation des données probantes énoncé par énoncé. Chaque affirmation institutionnelle cite sa source autorisée, produisant des réponses vérifiables et corrigibles au niveau de l'énoncé individuel plutôt que de la réponse entière.
- Un historique de déploiement et d'évaluation. L'architecture est opérationnelle dans plusieurs projets pilotes institutionnels (déploiements universitaires ; projet pilote d'un service fédéral) et a été soumise à des évaluations comparatives à l'aveugle contre les principaux systèmes de génération d'abord selon une matrice de critères publiée, incluant une pénalité d'hallucination et d'intrusion hors du champ autorisé, ainsi qu'à des tests répétés démontrant que son comportement à zéro source non autorisée est invariant d'une exécution à l'autre, là où les systèmes de génération d'abord varient.

L'élément probant issu des évaluations mérite une remarque de cadrage. Lors des tests comparatifs, les systèmes de génération d'abord ont parfois bien performé lors d'exécutions individuelles — y compris des exécutions sans aucune source non autorisée. Ce n'est pas un contre-exemple à la proposition architecturale ; c'est la proposition elle-même. Un bon comportement que rien n'impose est une disposition, observable exécution par exécution ; un comportement qu'une barrière d'exécution impose est une propriété, certifiable à travers les exécutions, les mises à jour de modèles, les invites et les utilisateurs. Dans l'IA de génération d'abord, la gouvernance est une invite. Dans l'IA de gouvernance d'abord, elle est une contrainte d'exécution. Les établissements peuvent vérifier des contraintes ; ils ne peuvent pas vérifier des dispositions.

7. Limites et questions ouvertes

Un positionnement honnête exige d'énoncer ce que le présent document n'établit pas. Premièrement, la portée : tel que déployé, COMPAiSS gouverne l'information institutionnelle publique ; il ne traite ni le

raisonnement clinique ou sur des données personnelles, ni l'autorisation d'actions agentiques (le territoire de CaMeL), ni l'alignement à l'entraînement, et il présuppose que les sources autorisées de l'établissement sont elles-mêmes exactes et tenues à jour — l'architecture garantit la provenance, non l'exactitude du contenu de cette provenance. Deuxièmement, l'échelle des évaluations : les évaluations comparatives menées à ce jour sont structurées mais modestes (quelques dizaines de questions, un petit nombre d'exécutions répétées), réalisées selon une matrice publiée et une notation à l'aveugle, mais non encore répliquées de façon indépendante ni révisées par les pairs. Troisièmement, le domaine lui-même : la quasi-totalité de la littérature la plus proche — y compris CaMeL, Llama Guard, Constitutional AI et les rapports de terminologie convergente — relève de la prépublication, de la documentation de fournisseurs ou de l'auto-publication ; le noyau révisé par les pairs de ce champ est encore en formation, et les positions prises ici devront être revues à mesure qu'il se constitue. Quatrièmement, les compromis de frontière : confiner l'inférence aux données probantes autorisées produit des réponses plus minces là où la couverture institutionnelle est mince ; nous considérons qu'il s'agit du bon compromis pour des établissements imputables, mais c'est un compromis, et la mesure de son coût pour la complétude des réponses est un travail en cours.

8. Conclusion

L'orientation du domaine ne fait plus sérieusement débat : le contrôle remonte en amont, du filtrage de ce que les modèles disent vers la gouvernance de ce qu'ils peuvent faire — et, dans la position développée ici, de ce à partir de quoi ils peuvent connaître une réponse. Au sein de ce mouvement, COMPAiSS occupe une position précise et vérifiable : c'est une implémentation d'une approche architecturale fondée sur l'autorisation préalable, qui se distingue par les données probantes autorisées par l'établissement avant l'inférence, une barrière d'exécution déterministe avec refus gouverné en mode fermé par défaut, la préservation des données probantes énoncé par énoncé et une base croissante d'évaluations en déploiement. La distinction qui compte pour des établissements imputables n'est pas de savoir quel modèle sous-jacent est le plus capable, mais quelle propriété de système ils peuvent certifier. Une contrainte d'exécution portant sur les données probantes est une telle propriété. Telle est la position que COMPAiSS revendique dans la littérature émergente, et la norme à l'aune de laquelle il invite à être évalué.

Références

Le statut de révision par les pairs est indiqué pour chaque entrée. Tous les registres ont été vérifiés comme existants et accessibles en juillet 2026 ; les titres sont cités dans leur langue originale. Un élément supplémentaire circulant dans des versions antérieures (un PDF hébergé sur arXiv intitulé « AI governance control stack ») est exclu ici, ses métadonnées bibliographiques stables n'ayant pu être vérifiées.

Akarkach, M. (2026a). Pre-inference governance – Canonical definition (version 1.1) [rapport technique, non révisé par les pairs]. Zenodo. https://doi.org/10.5281/zenodo.18401330

Akarkach, M. (2026b). Pre-inference governance beyond medicine: Authorization architectures for high-risk autonomous systems (version 1.0) [rapport technique, non révisé par les pairs]. Zenodo. https://doi.org/10.5281/zenodo.18261708

Akarkach, M. (2026c). Preventing epistemic exposure in frontier artificial intelligence systems: Pre-inference authorization as an architectural solution to inference extraction and discovery risk [rapport technique, non révisé par les pairs]. Zenodo. https://doi.org/10.5281/zenodo.18795429

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., … Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback [prépublication, non révisée par les pairs]. arXiv. https://doi.org/10.48550/arXiv.2212.08073

Debenedetti, E., Shumailov, I., Fan, T., Hayes, J., Carlini, N., Fabian, D., Kern, C., Shi, C., Terzis, A. et Tramèr, F. (2025). Defeating prompt injections by design [prépublication, non révisée par les pairs]. arXiv. https://doi.org/10.48550/arXiv.2503.18813

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., … Khabsa, M. (2023). Llama Guard: LLM-based input–output safeguard for human–AI conversations [prépublication, non révisée par les pairs]. arXiv. https://doi.org/10.48550/arXiv.2312.06674

NVIDIA Corporation. (2023–2026). NeMo Guardrails: Programmable guardrails for LLM applications ; NeMo Guardrails plugin for NeMo Relay: Runtime governance for LLM and tool execution [documentation technique, non révisée par les pairs]. https://docs.nvidia.com/nemo/guardrails/ ; https://github.com/NVIDIA-NeMo/Guardrails

COMPAiSS. (2026). Résultats d’évaluation comparative : questions-réponses Service Canada, protocole à l’aveugle, matrice de critères v3 ; et tests répétés de conformité à la gouvernance [rapports d’évaluation internes]. Offerts sur demande ; sommaires à https://compaiss.ai/demos