COUR SUPÉRIEURE DU QUÉBEC

RAPPORT D’ÉVALUATION

Projet pilote d’intelligence artificielle

Rapport complet • Avril 2026

Données collectées
Décembre 2025 – Mars 2026

SOMMAIRE EXÉCUTIF

La Cour supérieure du Québec a conduit, de décembre 2025 à mars 2026, un projet pilote visant à évaluer l'usage d'agents d'intelligence artificielle (IA) dans le cadre des fonctions judiciaires. Ce projet, mené sous l'autorité du bureau de la juge en chef, constitue à notre connaissance l'une des premières expériences documentées et contrôlées d'intégration de l'IA dans un tribunal supérieur canadien.

Vingt-deux (22) juges volontaires ont participé au projet, ce qui représente 11,17% du total des juges de la Cour supérieure. Ils étaient répartis en trois groupes selon un protocole expérimental en phases décalées. Neuf agents IA spécialisés ont été déployés via la plateforme Microsoft Teams (Copilot Studio), couvrant des fonctions d'aide à la rédaction, de traduction, de recherche législative, de citation juridique et de soutien technique.

Principaux résultats

89 % des répondants affirment que les limites des agents IA étaient compréhensibles et gérables. 84 % estiment que l'IA est compatible avec les exigences de la fonction judiciaire sous réserve d'un encadrement clair. 89 % se disent à l'aise de continuer à utiliser ce type d'outil dans un cadre balisé. 63 % recommandent un déploiement élargi. L'agent le plus utilisé (assistant rédacteur) a cumulé 479 sessions avec un taux d'engagement de 94 %.

Les agents se sont révélés particulièrement efficaces pour la révision et reformulation de texte (89 % des usages), la traduction (63 %), le soutien à la rédaction (63 %) et la recherche juridique (53 %). Les limites identifiées portent principalement sur la qualité variable des réponses en matière de recherche juridique spécialisée, la nécessité de vérification systématique, et l'absence d'interconnexion entre les agents.

Le rapport conclut que l'usage de l'IA est viable dans un environnement judiciaire sous réserve de conditions strictes d'encadrement, de formation et de supervision, et formule des recommandations concrètes pour la suite du projet.

1. INTRODUCTION ET CONTEXTE

1.1 Contexte institutionnel

La Cour supérieure du Québec est le tribunal de droit commun de première instance de la province. Elle est compétente en matière civile, commerciale, familiale, criminelle et en matière de droit public. Les juges de la Cour supérieure exercent leurs fonctions dans un environnement de travail exigeant, caractérisé par un volume élevé de dossiers, des exigences procédurales strictes et des tâches préparatoires nombreuses.

Dans ce contexte, l'émergence des technologies d'intelligence artificielle générative souligne la nécessité, pour les institutions judiciaires, d'explorer de manière rigoureuse et encadrée les possibilités offertes par ces outils. Les tribunaux canadiens et étrangers ont commencé à se saisir de cette question, mais peu ont mené des expériences structurées avec des juges en exercice.

1.2 Enjeux et motivation du projet

Les juges de la Cour supérieure ont exprimé, dans les années précédant le projet, un intérêt croissant pour les outils numériques susceptibles d'améliorer l'efficacité de leur travail préparatoire. Plusieurs utilisaient déjà, à titre personnel, des outils d'IA grand public (ChatGPT, Copilot, etc.) pour diverses tâches. Or, l'absence d'encadrement institutionnel posait des risques importants sur le plan de la confidentialité, de la fiabilité et de l'indépendance judiciaire.

Le projet pilote visait donc à offrir un cadre contrôlé permettant d'explorer ces outils en toute sécurité, tout en produisant des données empiriques sur leur efficacité et leur adéquation avec les exigences de la fonction judiciaire.

1.3 Objectifs du projet pilote

Le projet pilote poursuivait quatre objectifs principaux :

- Mesurer l'utilisation réelle des agents IA par les juges dans leurs fonctions quotidiennes.
- Évaluer la satisfaction et la perception d'utilité des agents déployés.
- Identifier les usages les plus pertinents et les limites des outils dans un contexte judiciaire.
- Définir un cadre de gouvernance reproductible pour un déploiement élargi potentiel.

2. CADRE DE GOUVERNANCE

2.1 Principes directeurs

Le cadre de gouvernance adopté par la Cour supérieure pour ce projet pilote[1] repose sur cinq principes fondamentaux :

- **Principe 1 :** Indépendance judiciaire

Aucun agent IA ne peut influer sur le jugement ou la décision du juge. L'utilisation est limitée aux tâches préparatoires et de soutien.

- **Principe 2 :** Confidentialité et sécurité des données

Les agents sont déployés au sein de l'infrastructure sécurisée Microsoft 365 de la Cour, garantissant que les données des dossiers ne transitent pas par des serveurs externes non contrôlés.

- **Principe 3 :** Transparence et vérification

Les juges sont tenus de vérifier toute information fournie par les agents avant utilisation dans leurs travaux. La réponse de l'IA est considérée comme un point de départ, non comme un résultat final.

- **Principe 4 :** Proportionnalité des usages

L'usage des agents est circonscrit à des catégories de tâches définies et approuvées, excluant notamment la rédaction finale de jugements.

- **Principe 5 :** Réversibilité et évaluation continue

Le projet est conçu pour permettre un arrêt à tout moment, et ses effets sont évalués de manière continue par des instruments méthodologiques robustes.

2.2 Encadrement de l'utilisation

Les agents IA déployés dans le cadre du projet sont conçus avec des paramètres prédéfinis limitant leur périmètre d'action. Chaque agent est spécialisé dans un domaine précis et ne peut répondre qu'aux requêtes liées à sa fonction. Cette architecture modulaire réduit le risque de dérives ou d'utilisations non anticipées.

Au-delà du périmètre thématique de chaque agent, des instructions explicites ont été intégrées dans les paramètres système afin de spécialiser non seulement le domaine d'intervention, mais également le format des résultats attendus. Chaque agent a ainsi été configuré pour produire exclusivement un type de réponse défini, correspondant à un besoin opérationnel précis du milieu judiciaire. Cette contrainte de format vise à assurer la fiabilité, la reproductibilité et l'utilité directe des extrants produits. Seulement deux agents, l'Assistant rédacteur et le Traducteur, ont reçu des instructions plus larges dans la mesure où les besoins en rédaction et en traduction peuvent difficilement se limiter à un seul format.

Cette approche de « spécialisation par extrant » constitue une mesure de gouvernance algorithmique concrète : elle réduit le risque d'hallucination interprétative, renforce la traçabilité des sources et s'aligne avec le principe d'indépendance judiciaire, en préservant entièrement le jugement du juge.

Une séance de formation a été offerte aux participants au début du projet, complétée par des rencontres périodiques hebdomadaires tenues en formule hybride (présentiel et Microsoft Teams) pour assurer un suivi de l'expérience et recueillir les commentaires des participants.

[1] Le cadre de gouvernance IA de la Cour supérieure du Québec est disponible à l'adresse suivante : https://coursuperieureduquebec.ca/fileadmin/cour-superieure/A_propos/2025-09_Cadre_gouvernance_IA_Cour_superieure_Quebec.pdf

2.3 Protection de la confidentialité

L'un des enjeux centraux identifiés dès la conception du projet était la protection de la confidentialité des dossiers judiciaires. Les sondages ont révélé que la confidentialité et la sécurité des données figuraient parmi les préoccupations les plus fréquemment mentionnées par les participants (sondage de départ). La solution retenue, basée sur Microsoft 365 et Copilot Studio, s'inscrit dans un protocole établi par le ministère de la Justice du Québec (MJQ) concernant l'hébergement et le traitement des données judiciaires.

Par ailleurs, une décision architecturale importante a été prise : plutôt qu'un outil s'abreuvant automatiquement aux documents présents sur les appareils des utilisateurs, les agents ont été conçus selon un modèle « bac à sable » que chaque juge devait alimenter manuellement. L'accès à chaque agent était réservé personnellement au juge concerné, sans partage de sessions ni d'historiques entre utilisateurs.

Malgré l'ensemble de ces mesures, il n'était pas possible de garantir un traitement à cent pour cent de l'information en sol canadien. En conséquence, les juges ont été expressément instruits d'éviter de soumettre aux agents tout document ou texte contenant des renseignements personnels ou des informations confidentielles relatives aux dossiers en cours.

Avantage clé

Plusieurs participants ont spécifiquement mentionné l'environnement sécurisé comme principal avantage différenciateur des agents du projet par rapport aux outils d'IA commerciaux grand public, permettant d'exploiter des fonctionnalités similaires dans un cadre institutionnellement fiable.

3. MÉTHODOLOGIE

3.1 Design expérimental

Le projet pilote a été officiellement déployé le 8 décembre 2025 et s'est conclu le 16 mars 2026, pour une durée totale d'expérimentation de 98 jours (environ 14 semaines). Cette période inclut la phase d'accès aux agents, la collecte de données et l'administration du sondage final.

Le projet a adopté un design expérimental en groupes décalés, inspiré des protocoles de type « stepped-wedge » utilisés dans la recherche en santé et en sciences sociales. Ce design présente l'avantage de permettre une comparaison intragroupe (avant/après accès aux outils) tout en limitant les effets de sélection. Il permet également d'observer l'effet du retrait de l'accès sur la perception de la valeur de l'outil.

Trois groupes ont été constitués :

Groupe	Modalité d'accès	N (sondage final)
Groupe A	Accès dès le début, pendant toute la durée	7 (37 %)
Groupe B	Accès retiré avant la fin du projet	6 (32 %)
Groupe C	Accès débutant plus tard dans le projet	6 (32 %)

3.2 Sélection et profil des participants

La participation au projet était volontaire. Quarante-deux (42) juges de la Cour supérieure ont manifesté leur intérêt. Parmi eux, 22 juges ont été sélectionnés pour constituer le groupe pilote selon une procédure en deux étapes : les candidats ont d'abord été regroupés en sous-groupes définis par des critères de représentativité (diversité des matières : civil, criminel, famille, commercial, mixte ; distribution géographique : Montréal, Laval, Saint-Maurice et autres districts ; niveau d'aisance technologique préalable), puis une sélection aléatoire a été effectuée au sein de chaque sous-groupe. Cette démarche visait à constituer un échantillon aussi représentatif que possible de l'ensemble du corps judiciaire de la Cour supérieure, tout en évitant les biais liés à l'auto-sélection volontaire.

Initialement envisagé, un groupe témoin n'a finalement pas été constitué. L'intensité de l'accompagnement requis (rencontres hebdomadaires, suivi individualisé et animation continue) rendait peu réaliste la gestion simultanée d'un groupe contrôle avec la même rigueur. Le design expérimental a donc été adapté pour s'appuyer sur les comparaisons avant/après propres à chaque participant.

Le profil des répondants au sondage final (n = 19) est le suivant :

- Matières : Mixte/plusieurs matières (47 %), Civil (26 %), Famille (11 %), Criminel (11 %), Autre (5 %)
- Aisance technologique préalable : Moyenne (42 %), Plutôt élevée (32 %), Très élevée (21 %), Plutôt faible (5 %)

- Usage d'IA avant le projet (sondage de départ, n=13) : Oui à l'occasion (38 %), Oui régulièrement (31 %), Rarement (23 %), Jamais (8 %)

3.3 Instruments de mesure

Cinq instruments de collecte de données ont été utilisés, permettant une triangulation des résultats :

- **Instrument 1 :** Sondage de départ pour le groupe déployé tardivement (n=6, décembre 2025) : administré avant le début du projet à un sous-groupe de participants ayant intégré l'expérimentation tardivement, ce sondage mesurait l'aisance technologique de base, les usages antérieurs de l'IA, les préoccupations initiales et les attentes. Les résultats indiquent que 100 % des répondants identifiaient la confidentialité comme principale préoccupation, 83 % la sécurité des données, que 5 sur 6 étaient à l'aise avec Microsoft Teams et que 4 sur 6 avaient une expérience antérieure de l'IA.
- **Instrument 2 :** Sondage de départ principal (n=13, décembre 2025) : mesure des connaissances préalables, attentes, niveaux d'aisance technologique, et préoccupations initiales.
- **Instrument 3 :** Sondage d'évaluation à deux mois (n=17, février 2026) : mesure de l'adoption initiale, des usages, et satisfaction après première période d'utilisation.
- **Instrument 4 :** Données d'utilisation : métriques quantitatives collectées automatiquement par la plateforme (sessions, engagement, qualité des réponses, réactions des utilisateurs).
- **Instrument 5 :** Sondage final (n=19, mars 2026) : évaluation globale de l'expérience, mesure de l'impact perçu, recommandations pour la suite. Il a été administré à la toute fin du projet.

Le sondage final comportait 42 questions, dont 14 items sur échelle de Likert à 5 points, 8 questions à choix multiples, une échelle de recommandation de 0 à 10 (Net Promoter Score adapté (NPS)), et des questions ouvertes recueillant les verbatims des participants.

3.4 Évaluation de la qualité des agents : questions synthétiques

En complément des sondages, une évaluation structurée de la qualité des réponses des agents a été réalisée à l'aide de questions synthétiques préparées à l'interne. Un total de 90 évaluations a été mené sur 9 agents spécialisés (10 questions par agent), selon une méthodologie différenciée selon la nature de l'agent.

Pour les agents juridiques à base documentaire (Code civil, Code de procédure civile, Code criminel, Loi sur la faillite), chaque évaluation comportait trois critères : (1) citation verbatim de l'article attendu (oui/non), (2) pertinence (faible, moyenne, élevée), et (3) complétude de la réponse (incomplète, partielle, complète), avec un score synthétique sur 10. Pour l'agent Traducteur, six critères spécialisés ont été évalués (fidélité, fluidité, registre, cohérence terminologique, structure et adaptation culturelle). Pour l'Assistant rédacteur, les critères étaient la grammaire, la syntaxe, le vocabulaire juridique et la fidélité au sens. Le Citateur a fait l'objet d'une évaluation partielle portant sur la correction de citations selon le guide interne du Service de recherche de la Cour. Pour tous les agents, une relance était possible en cas de réponse initiale incomplète.

3.5 Accompagnement hebdomadaire et émergence des cas d'usage

Tout au long du projet pilote, les participants ont bénéficié d'un suivi structuré sur une base hebdomadaire. Des rencontres semi-dirigées régulières réunissaient les juges participants afin d'assurer une progression collective dans l'appropriation des outils.

3.5.1 Format et objectifs des rencontres

Chaque rencontre hebdomadaire poursuivait trois objectifs complémentaires : (1) une composante de sensibilisation portant sur les avantages et les risques associés à l'IA générative, adaptée aux enjeux propres au contexte judiciaire (hallucinations, biais, indépendance, confidentialité) ; (2) une composante de démonstration pratique axée sur des cas d'usage concrets identifiés au fil des tests d'utilisateurs ; et (3) un espace d'échange permettant aux juges de partager leurs expériences, leurs difficultés et leurs découvertes.

Ce format semi-dirigé a favorisé une dynamique d'apprentissage par les pairs : les cas d'usage qui émergeaient des pratiques d'un juge étaient rapidement partagés et testés par l'ensemble du groupe. Cette boucle de rétroaction rapide entre usage réel et formation a été un facteur important d'enrichissement du projet.

3.5.2 Cas d'usage émergents

L'exploration collective au fil des semaines a permis de découvrir et de valider des cas d'usage qui n'avaient pas été initialement anticipés dans la conception du projet. Parmi les plus significatifs :

Cas d'usage	Description et valeur ajoutée
Reconnaissance de notes manuscrites numérisées	Les agents ont été mis à contribution pour transcrire et structurer des notes écrites à la main et numérisées (scans), réduisant considérablement le temps de retranscription manuelle pour les juges.
Transformation de texte en présentation PowerPoint	La capacité de convertir automatiquement un texte narratif en diapositives structurées a été identifiée comme utile pour la préparation de présentations institutionnelles et pédagogiques.
Création de tableaux résumés avec passages et références intégrés	Les agents ont été utilisés pour générer des tableaux synthétiques incorporant des extraits de textes juridiques avec leurs références, facilitant la comparaison de sources ou la synthèse d'autorités.
Dépannage par capture d'écran (agent Technicien)	L'agent Technicien informatique IA a été utilisé en lui soumettant directement des captures d'écran de problèmes techniques, lui permettant d'analyser visuellement la situation et de proposer des solutions sans intervention du service TI.

Ces cas d'usage émergents illustrent l'importance d'un accompagnement structuré dans tout projet d'intégration d'IA : les utilisateurs découvrent progressivement des possibilités d'usage qui dépassent les cas prévus initialement, à condition de disposer d'un espace sécurisé pour expérimenter et partager. La régularité des rencontres a également permis d'ajuster le paramétrage de certains agents en cours de projet pour mieux répondre aux besoins observés.

Apport méthodologique

L'accompagnement hebdomadaire constitue, à notre connaissance, un élément distinctif de ce projet pilote par rapport à d'autres expérimentations d'IA dans le secteur public. Il a permis de combiner une évaluation rigoureuse avec une démarche d'apprentissage organisationnel, tout en assurant que l'usage des outils évolue dans un cadre éthique et institutionnellement balisé.

4. AGENTS IA DÉPLOYÉS

4.1 Architecture générale

Les agents IA ont été déployés via Microsoft Copilot Studio intégré à Microsoft Teams. Cette architecture présente plusieurs avantages : intégration native à l'outil de collaboration déjà utilisé par les juges, isolation des données dans l'environnement Microsoft 365 du gouvernement du Québec, et capacité de déployer des agents spécialisés avec des connaissances intégrées (documents juridiques, codes, etc.).

4.2 Description des agents

Agent	Fonction principale
Assistant rédacteur	Aide à la rédaction, reformulation, amélioration stylistique de textes juridiques et autres documents préparatoires.
Traducteur	Traduction français-anglais et anglais-français de passages juridiques, avec attention aux équivalences terminologiques.
Code civil du Québec IA	Recherche et explication d'articles du Code civil du Québec, avec citation des dispositions pertinentes.
Code de procédure civile IA	Recherche et explication d'articles du Code de procédure civile, avec identification des règles applicables.
Code criminel IA	Recherche dans le Code criminel, identification des infractions et peines applicables.
Loi sur la faillite IA	Recherche spécialisée dans la Loi sur la faillite et l'insolvabilité.
Le citateur	Correction et restructuration des références de sources juridiques.
Technicien informatique IA	Soutien technique pour les problèmes liés à l'utilisation des outils numériques de la Cour.
Livre bleu 2.0	Reproduction verbatim des sections pertinentes de cet ouvrage de doctrine interne en droit de la famille, préparé à l'usage exclusif des juges. L'agent ne procédait à aucune synthèse ni réinterprétation du contenu.

4.3 Paramétrage des agents

Le paramétrage des agents constitue un élément central de la fiabilité et de la sécurité du projet pilote. Chaque agent a été configuré dans Microsoft Copilot Studio selon une architecture commune assortie de personnalisations propres à chaque cas d'usage.

4.3.1 Environnement technologique et résidence des données

Les agents ont été déployés dans l'environnement Microsoft 365 du ministère de la Justice du Québec (MJQ), réservé aux juges. Cet environnement offre : (1) le chiffrement de bout en bout de toutes les communications, (2) l'hébergement des données visant le territoire canadien, (3) l'isolation des agents dans un modèle de type « bac à sable » sans accès automatique aux documents des appareils des utilisateurs, et (4) l'intégration native à Microsoft Teams, déjà utilisé par les juges. L'environnement a été validé préalablement au lancement du projet par des experts en cybersécurité qui conseillent les tribunaux québécois.

4.3.2 Bases de connaissances et instructions système

Chaque agent a été alimenté par une base de connaissances spécifique : les agents juridiques intègrent les textes légaux complets (Code civil du Québec, Code de procédure civile, Code criminel, Loi sur la faillite et l'insolvabilité), le Livre bleu (droit de la famille) de la Cour supérieure et des documents internes préparés par les juges et le personnel de soutien. Ces documents ont été spécialement mis en page pour optimiser leur traitement par l'IA, notamment avec des scripts dans Microsoft Word pour faciliter la reconnaissance des notes de bas de page en les intégrant directement dans le texte au long. Des instructions système (« system prompts ») détaillées ont été rédigées pour les agents afin d'orienter le style de réponse, imposer la citation verbatim des sources, définir le périmètre des sujets admissibles et formaliser les comportements attendus en cas de demande hors périmètre.

Cette intégration est effectuée exclusivement à des fins internes, dans un objectif de repérage et de soutien à la recherche juridique, et s'inscrit directement dans l'exercice des fonctions judiciaires.

Dans ce contexte, l'utilisation des textes législatifs ne constitue pas une reproduction à des fins de diffusion ou de commercialisation, mais une utilisation fonctionnelle inhérente à la mission du tribunal.

Le tribunal demeure par ailleurs tributaire des versions officielles des lois, et les outils développés ne sauraient s'y substituer.

De plus, la navigation sur le web a été désactivée pour chacun des agents, et l'accès à leur base de connaissances par défaut (la connaissance générale intégrée au modèle sous-jacent) a également été bloqué. Les agents étaient ainsi redirigés exclusivement vers les bases de connaissances désignées, garantissant que leurs réponses s'appuient uniquement sur les sources vérifiées et validées par l'équipe de coordination.

4.3.3 Restrictions intégrées et garde-fous

Conformément au cadre de gouvernance, des restrictions ont été intégrées directement dans la configuration de chaque agent. La capacité de génération et la longueur de la fenêtre de contexte ont été volontairement limitées afin de réduire le risque d'hallucinations. Les agents sont configurés pour refuser automatiquement les demandes portant sur : l'assistance au processus décisionnel judiciaire, la génération de projets de jugements, l'interprétation de textes juridiques, le résumé de textes volumineux et la formulation d'opinions juridiques. Ces garde-fous assurent que les agents demeurent des outils de soutien et non des substituts au raisonnement judiciaire. Chaque session s'ouvre par un avertissement rappelant à l'utilisateur la nécessité de vérifier tout contenu généré.

L'équipe est consciente des limites inhérentes aux instructions système : elles constituent une couche de protection supplémentaire, mais ne sont pas infaillibles face à des utilisateurs motivés. Elles s'inscrivent dans une approche de défense en profondeur combinée avec la formation des participants, la sensibilisation aux risques et les obligations professionnelles propres à la fonction judiciaire.

5. RÉSULTATS

5.1 Données d’utilisation

Les données collectées automatiquement par la plateforme Microsoft Copilot Studio sur la durée du projet permettent de brosser un portrait quantitatif de l’adoption des agents. Le tableau suivant présente les indicateurs clés par agent. À noter que certains indicateurs de satisfaction liés à certains agents n’ont pas pu être mesurés en raison du faible taux de rétroaction des utilisateurs.

Agent	Sessions	Croissance	Engagement	Satisfaction	Observation clé
Assistant rédacteur	479	+565 %	94 %	76,9 %	Agent le plus utilisé. Forte demande.
Traducteur	154	+267 %	94 %	89,6 %	286 questions posées. Forte croissance tout au long du projet.
Code de procédure civile	67	+49 %	96 %	N.D.	78 questions. Croissance stable et régulière.
Citateur	33	+74 %	79 %	90,9 %	Usage en hausse marquée sur la période.
Technicien IA	40	+25 %	85 %	90,9 %	Usage croissant pour le dépannage et les impressions d’écran.
Code civil du Québec	54	+116 %	85 %	N.D.	Usage doublé. 65 questions posées. Score synthétique : 9,0/10.
Livre bleu 2.0	38	+36 %	87 %	N.D.	41 % des répondants l’ont utilisé (sondage à deux mois). Citation verbatim parfaite.
Code criminel	17	−15 %	41 %	N.D.	Légère baisse. Questions complexes. Score 7,8/10 avec forte variabilité.
Loi sur la faillite	11	−45 %	45 %	N.D.	Usage très spécialisé. Score synthétique : 9,0/10.

Au total, 893 sessions ont été enregistrées sur la plateforme durant la période d’évaluation de 90 jours, tous agents confondus. L’assistant rédacteur domine largement avec 479 sessions, soit près de trois fois plus que le second agent (le traducteur, 154 sessions). Cette dominance reflète l’intérêt des juges pour les outils de reformulation, révision et amélioration stylistique. Le traducteur affiche la croissance la plus forte (+267 %), témoignant d’une appropriation progressive et soutenue tout au long du projet. Les agents de recherche juridique (Code civil, Code de procédure civile, Code criminel, Loi sur la faillite) ont également été adoptés, bien qu’avec des volumes plus modérés conformément à leur nature spécialisée. La croissance de l’ensemble des agents est positive, ce qui suggère une adoption cumulative plutôt qu’un effet de nouveauté.

5.2 Résultats des sondages

5.2.1 Fréquence et contextes d'utilisation

Le sondage final (n=19) révèle que la grande majorité des participants a adopté les agents IA dans le cadre de leur pratique judiciaire, avec une fréquence significative :

- Quelques fois par semaine : 63 % (12/19)
- Presque chaque jour de travail : 16 % (3/19)
- Environ une fois par semaine : 16 % (3/19)
- Moins d'une fois par semaine : 5 % (1/19)

Ces résultats indiquent une adoption robuste pour un projet pilote de première génération, avec 79 % des participants utilisant les agents au moins plusieurs fois par semaine.

Les contextes d'utilisation les plus fréquents sont, par ordre décroissant :

- Révision ou reformulation de texte : 89 % (17/19)
- Traduction ou soutien linguistique : 63 % (12/19)
- Soutien à la rédaction : 63 % (12/19)
- Recherche juridique ou doctrinale : 53 % (10/19)
- Repérage législatif ou réglementaire : 42 % (8/19)
- Résumé de textes ou de documents : 32 % (6/19)

Ces données confirment que la valeur perçue est la plus forte pour les tâches linguistiques et rédactionnelles, et plus modérée pour la recherche juridique spécialisée.

5.2.2 Résultats de l'échelle de Likert (sondage final)

Le tableau suivant présente les résultats des 14 items de l'échelle de Likert du sondage final (n=19) :

Énoncé	Accord*	Neutre	Désaccord
Les agents IA m'ont aidé à accomplir certaines tâches plus efficacement	**84 %**	5 %	11 %
Les agents IA m'ont permis d'explorer des pistes de recherche plus rapidement	**47 %**	26 %	26 %
Les agents IA ont contribué à améliorer la qualité de mes extrants préparatoires	**58 %**	26 %	16 %
Les agents IA se sont avérés utiles pour mes tâches préparatoires et répétitives	**47 %**	37 %	16 %
Les agents IA se sont bien intégrés à mon quotidien de travail	**68 %**	21 %	11 %
Les agents IA m'ont été utiles sans nuire à mon autonomie professionnelle ni à mon jugement	**84 %**	11 %	5 %
Les agents IA ont représenté un apport concret dans ma pratique	**63 %**	26 %	11 %
Les réponses des agents IA sont suffisamment fiables pour un usage préparatoire	**74 %**	16 %	11 %
Une vérification importante est requise avant d'utiliser les réponses des agents IA dans mon travail	**47 %**	32 %	21 %

Énoncé	Accord*	Neutre	Désaccord
Les agents IA m'ont parfois orienté vers des pistes erronées ou trompeuses	**21 %**	37 %	42 %
Les limites des agents IA sont compréhensibles et gérables dans mon contexte de travail	**89 %**	5 %	5 %
Le cadre d'utilisation établi pour le projet était adéquat	**74 %**	0 %	26 %
Les agents IA sont compatibles avec les exigences de la fonction judiciaire, pourvu qu'un encadrement clair soit en place	**84 %**	16 %	0 %
Je serais à l'aise de continuer à utiliser ce type d'outil dans un cadre bien balisé	**89 %**	0 %	11 %

** Accord = « Plutôt d'accord » + « Tout à fait d'accord ». Désaccord = « Plutôt en désaccord » + « Pas du tout d'accord ».*

5.2.3 Recommandation et perception globale

Le sondage final a également permis de recueillir les intentions et recommandations des participants quant à la suite du projet :

- 63 % (12/19) recommandent un déploiement plus large de l'outil à l'ensemble des juges
- 16 % (3/19) recommandent un élargissement graduel à d'autres juges volontaires
- 16 % (3/19) sont incertains quant à la suite
- 5 % (1/19) recommandent une prolongation de l'expérimentation à petite échelle

> **Indicateur clé**
> En additionnant les 63 % favorables à un déploiement élargi et les 16 % favorables à un élargissement graduel, on constate que 79 % des participants (15 sur 19) souhaitent une prolongation et un élargissement du recours à l'IA dans leurs pratiques judiciaires. Ce chiffre constitue un signal fort en faveur de la pérennisation du programme, sous réserve de l'amélioration continue des agents et du maintien d'un cadre rigoureux.

Quant à la recommandation de poursuite globale, 68 % (13/19) des répondants la recommandent sans réserve, 26 % (5/19) sont incertains et 5 % (1/19) ne la recommandent pas. Cette distribution, combinée au taux de 79 % favorable à une forme d'élargissement, suggère que même les participants plus hésitants reconnaissent la valeur du projet et envisagent sa suite, tout en souhaitant une approche plus progressive.

5.3 Évaluation à deux mois (n = 17)

Le sondage d'évaluation à deux mois a permis de mesurer l'expérience des participants après une première période d'utilisation des agents dans Teams. Les principaux résultats sont les suivants :

- Satisfaction globale : 47 % satisfaits, 29 % moyennement satisfaits, 18 % peu satisfaits, 6 % très satisfaits
- Gain de temps : 29 % beaucoup, 47 % modérément, 12 % un peu ou très peu
- Intégration aux habitudes de travail : 35 % très facilement, 29 % facilement, 24 % moyennement, 12 % difficilement

- Souhait de continuer après la fin du projet : 76 % « Oui certainement », 12 % « Probablement », 12 % « Peut-être »
- Agents les plus utilisés : Assistant rédacteur (100 %), Traducteur (71 %), Lois IA (53 %), Livre bleu (41 %)

Les principaux facteurs contribuant à la satisfaction sont la facilité d'utilisation (65 % des répondants), la rapidité (53 %) et la qualité des réponses (47 %). Les principaux facteurs limitants sont la qualité variable des réponses (47 %) et le manque de précision juridique (47 %).

5.4 Comparaison avec le Compagnon CAIJ

Dans le cadre du sondage à deux mois, les participants ont été invités à comparer les agents IA du projet pilote avec le Compagnon CAIJ, outil d'IA développé par le Centre d'accès à l'information juridique, auquel ils avaient accès en mode bêta.

- Fiabilité des résultats pour la recherche juridique : 71 % estiment le CAIJ supérieur, 18 % les jugent comparables
- Rapidité et efficacité : 41 % CAIJ supérieur, 29 % comparables, 24 % agents IA supérieurs
- Facilité d'intégration aux habitudes : 35 % agents IA supérieurs, 35 % comparables
- Préférence pour la recherche juridique : 71 % CAIJ, 18 % les deux selon contexte

> **Interprétation**
> Ces résultats suggèrent une complémentarité plutôt qu'une concurrence entre les deux outils, notamment en ce qui concerne la rapidité et l'intégration à Teams. Le Compagnon CAIJ est perçu comme plus spécialisé pour la recherche juridique, tandis que les agents du projet pilote sont appréciés pour leur facilité d'intégration à Teams et leur polyvalence (rédaction, traduction, soutien technique).

5.5 Résultats des questions synthétiques par agent

L'évaluation par questions synthétiques a porté sur l'ensemble des 9 agents déployés. Selon la nature de l'agent, deux grilles d'évaluation ont été employées : une grille standard (citation verbatim, pertinence, complétude, score /10) pour les agents référençant des textes sources, et des grilles spécialisées pour le Traducteur, l'Assistant rédacteur et le Citateur.

5.5.1 Agents à base documentaire (n = 60 questions)

Les six agents référençant directement des textes légaux ou doctrinaux ont été soumis à 10 questions chacun, évaluées sur trois critères (citation verbatim, pertinence, complétude) et un score global /10.

Agent	Questions	Score moy.	Citation verbatim	Pertinence élevée	Complétude
Code civil du Québec	10	**9,0 / 10**	100 %	90 %	70 %
Loi sur la faillite	10	**9,0 / 10**	100 %	80 %	80 %
Code de procédure civile	10	**8,9 / 10**	100 %	70 %	70 %
Technicien	10	**8,6 / 10**	N/A	90 %	70 %

Agent	Questions	Score moy.	Citation verbatim	Pertinence élevée	Complétude
informatique					
Livre bleu 2.0 (droit de la famille)	10	**7,0 / 10**	100 %	90 %	40 %
Code criminel	10	**7,8 / 10**	100 %	80 %	50 %

Les agents Code civil du Québec et Loi sur la faillite obtiennent les meilleurs scores globaux (9,0/10), avec une citation verbatim parfaite. La Loi sur la faillite affiche également le taux de complétude le plus élevé parmi les agents à base documentaire (80 %). Le Code de procédure civile (8,9/10) maintient une citation verbatim à 100 %, mais une complétude de 70 %, reflétant des réponses parfois partielles sur des questions procédurales complexes. Le Technicien informatique (8,6/10), qui ne cite pas de texte source, affiche 90 % de pertinence élevée et 70 % de complétude. L'agent Livre bleu 2.0 (7,0/10) obtient une citation verbatim parfaite (100 %) et une pertinence élevée dans 90 % des cas, mais son taux de complétude (40 %) est le plus faible du groupe, indiquant que les réponses couvrent correctement les notions de base en droit de la famille mais manquent fréquemment d'une structure analytique complète. L'agent Code criminel (7,8/10) présente la plus grande variabilité (plage 3/10–9/10) et un taux de complétude de 50 %, reflétant la complexité et l'ambiguïté fréquentes des questions de droit criminel.

5.5.2 Agents spécialisés avec grilles d'évaluation propres

L'agent Traducteur (n = 10 textes) a été évalué selon six critères spécialisés :

Critère d'évaluation (Traducteur)	Score moyen /10	Score min.
Fidélité au sens	**10,0 / 10**	10,0
Registre et ton	**10,0 / 10**	10,0
Structure et ponctuation	**9,9 / 10**	9,0
Adaptation culturelle	**9,8 / 10**	9,0
Fluidité	**9,6 / 10**	9,0
Cohérence terminologique	**9,5 / 10**	8,0

L'agent Traducteur présente des résultats excellents : la fidélité au sens et le registre juridique obtiennent un score parfait de 10/10 sur l'ensemble des textes évalués. La cohérence terminologique constitue le critère le plus variable (min. 8,0/10), ce qui reflète la complexité des équivalences entre les termes juridiques français et anglais.

L'Agent rédacteur (n = 10 textes) a été évalué selon quatre critères : Grammaire (9,6/10), Fidélité au sens (9,8/10), Vocabulaire juridique (9,0/10), Syntaxe (8,9/10). Les performances les plus élevées sont constatées sur les textes aux enjeux stylistiques simples. Les textes présentant des ellipses syntaxiques ou des termes techniques spécialisés donnent lieu à des résultats légèrement plus variables (min. 7,0/10 pour le vocabulaire juridique).

Le Citateur (n = 10 questions évaluées) obtient un score moyen de 8,8/10, avec une pertinence élevée à 100 % et une complétude de 90 %. L'agent maîtrise très bien la correction formelle des citations selon le guide interne du Service de recherche de la Cour, produisant des résultats parfaits

(10/10) dans 9 cas sur 10. La seule exception notable concerne une question portant sur un article de revue : l'agent a inventé le prénom de l'autrice, constituant une hallucination factuelle qui a entraîné un score de 0/10. Cet incident isolé est significatif : bien que l'agent soit très fiable pour les types de citations qu'il maîtrise, il peut échouer de façon inattendue pour les sources doctrinales moins standardisées. Au-delà de cette limite, les résultats confirment la fiabilité de l'agent dans son périmètre actuel, tout en rappelant que cette spécialisation à la correction formelle en limite l'utilité opérationnelle. L'agent gagnerait à être élargi par l'intégration d'une banque des sources fréquemment citées devant la Cour supérieure, permettant à terme la vérification du fond et la recherche des références complètes.

Indicateur	Résultat
Questions évaluées	10 questions (sur 11 posées)
Score moyen	8,8 / 10
Questions parfaites (10/10)	9 questions sur 10
Pertinence élevée	100 %
Complétude complète	90 %
Hallucination factuelle détectée	**1 cas — prénom d'autrice inventé (score : 0/10)**

5.5.3 Synthèse comparative et lecture transversale

L'évaluation complète des 9 agents permet d'établir un classement de performance clair. Les agents de rédaction et de traduction dominent : le Traducteur (9,8/10) et l'Assistant rédacteur (9,3/10) offrent les résultats les plus constants et les plus élevés. Le Citateur (8,8/10) s'inscrit juste en dessous, avec d'excellentes performances sur les citations standardisées mais une vulnérabilité aux hallucinations factuelles pour les sources moins répertoriées. Les agents à base documentaire juridique se situent dans la fourchette 8,6–9,0/10, à l'exception du Code criminel (7,8/10), dont la variabilité marquée (3/10–9/10) traduit la complexité propre au droit criminel. Le Livre bleu 2.0 (7,0/10) constitue un cas particulièrement instructif : malgré une citation verbatim parfaite, son taux de complétude (40 %) est le plus bas de l'ensemble, suggérant que l'agent reproduit fidèlement les passages pertinents sans toujours couvrir la totalité de la question posée, ce qui traduit un déficit de structure analytique plutôt que de contenu. Ces résultats sont cohérents avec les verbatims des sondages, qui identifient la rédaction et la traduction comme les usages les plus satisfaisants, et la recherche juridique spécialisée comme le domaine nécessitant le plus d'amélioration.

Observation méthodologique

Les questions synthétiques représentent des scénarios idéalisés : les questions ont été formulées de façon à tester des compétences spécifiques. Elles sont également évaluées à l'interne et selon les attentes subjectives de leur auteur. Dans la pratique, les juges posent des questions plus complexes, plus contextuelles et souvent multidimensionnelles, ce qui peut engendrer une performance réelle inférieure aux scores observés dans ce cadre structuré. Il demeure néanmoins que les agents arrivaient à produire le texte verbatim des dispositions applicables, c'est au niveau de la complétude et de la pertinence que les points étaient généralement perdus.

5.6 Analyse complémentaire : matrice performance synthétique / satisfaction perçue

La mise en regard des scores d'évaluation synthétique (Annexe C) et des indicateurs de satisfaction perçue issus des données de plateforme et des sondages permet d'identifier le positionnement relatif de chaque agent selon deux axes indépendants : la performance mesurée objectivement et la valeur perçue par les utilisateurs. Cette analyse permet de distinguer les agents surestimés ou sous-estimés de ceux dont la perception est alignée avec la performance réelle.

Agent	Score éval. /10	Satisfaction plateforme	Positionnement	Observation analytique
Traducteur	**9,8 / 10**	89,6 %	*Perf. élevée — Satisfaction élevée*	Alignement optimal. Agent dont la valeur perçue est cohérente avec l'excellence mesurée.
Assistant rédacteur	**9,3 / 10**	76,9 %	*Perf. élevée — Satisfaction modérée*	Agent sous-estimé. Malgré le meilleur score composite et le volume d'usage le plus élevé (479 sessions), la satisfaction plateforme est la plus basse. Cela peut refléter des attentes plus élevées ou des usages plus complexes.
Code civil du Québec	**9,0 / 10**	N.D.	*Perf. élevée — Satisfaction non mesurée*	Excellent score synthétique (citation verbatim 100 %, pertinence 90 %). Absence de donnée de satisfaction par agent sur la plateforme.
Loi sur la faillite	**9,0 / 10**	N.D.	*Perf. élevée — Satisfaction non mesurée*	Score comparable au CCQ avec le meilleur taux de complétude de son groupe (80 %).
Code de procédure civile	**8,9 / 10**	N.D.	*Perf. élevée — Satisfaction non mesurée*	Excellent score synthétique. Absence de donnée de satisfaction directe par agent sur la plateforme.
Citateur	**8,8 / 10**	90,9 %	*Perf. élevée — Satisfaction élevée*	Alignement positif. Satisfaction élevée malgré l'hallucination détectée, suggérant que les utilisateurs valorisent la fiabilité générale sans percevoir nécessairement les cas limites.
Technicien IA	**8,6 / 10**	90,9 %	*Perf. correcte — Satisfaction élevée*	Léger surclassement perceptif. La satisfaction est élevée relativement au score évalué, ce qui peut s'expliquer par la nature pratique et immédiate du soutien technique offert.

Agent	Score éval. /10	Satisfaction plateforme	Positionnement	Observation analytique
Code criminel	**7,8 / 10**	N.D.	*Perf. modérée — Satisfaction non mesurée*	Variabilité marquée (plage 3/10–9/10), cohérente avec la complexité des questions de droit criminel.
Livre bleu 2.0	**7,0 / 10**	N.D.	*Perf. modérée — Satisfaction non mesurée*	Score le plus bas du groupe. Malgré une citation verbatim parfaite, le taux de complétude (40 %) limite la valeur opérationnelle. Usage tout de même notable (41 % au sondage à deux mois).

** Satisfaction plateforme : taux de recommandation (pouce haut) issu des données Microsoft Copilot Studio. Engagement : taux de sessions ayant généré une réponse utilisateur. N.D. : indicateur non disponible à ce niveau de granularité.*

Limite méthodologique : fatigue de rétroaction

Les taux de satisfaction issus de la plateforme Microsoft Copilot Studio reposent sur les réactions volontaires des utilisateurs (pouce haut/bas). Il est possible qu'une fatigue de rétroaction se soit installée au fil du temps, les participants les plus assidus ayant pu réduire leurs évaluations manuelles pour les sessions routinières. Ces chiffres doivent donc être interprétés comme des indicateurs tendanciels plutôt que comme des mesures absolues de satisfaction.

Interprétation : cas de l'Assistant rédacteur

L'Assistant rédacteur constitue le cas analytique le plus instructif : il présente le score d'évaluation synthétique le plus élevé après le Traducteur (9,3/10), le volume d'usage de loin le plus important (479 sessions), mais la satisfaction plateforme la plus basse (76,9 %). Cette tension entre performance mesurée et valeur perçue peut s'expliquer par deux hypothèses complémentaires : (1) des attentes utilisateurs plus élevées pour un agent généraliste et fréquemment sollicité, et (2) une utilisation pour des tâches plus complexes et contextuelles, où les limites du modèle se manifestent davantage que pour des tâches structurées comme la traduction ou le soutien technique. Ce phénomène est cohérent avec la littérature sur l'adoption des outils IA : les outils les plus utilisés sont aussi ceux sur lesquels les utilisateurs projettent les attentes les plus diversifiées.

5.7 Analyse comparative par groupe expérimental

Le design en groupes décalés permet de comparer l'expérience des participants selon leur modalité d'accès : accès complet sur toute la durée (Groupe A, n = 7), accès retiré avant la fin du projet (Groupe B, n = 6) et accès débutant plus tardivement (Groupe C, n = 6). Cette analyse repose sur les réponses au sondage final (n = 19), les participants ayant été identifiés par leur groupe via une question dédiée du sondage. Les 3 participants non-répondants ne sont pas attribués à un groupe spécifique.

5.7.1 Indicateurs clés par groupe

Indicateur	Groupe A (accès complet)	Groupe B (accès retiré)	Groupe C (accès tardif)
N (sondage final)	7	6	6
Fréquence élevée (≥ quelques fois/sem.)	86 %	67 %	83 %
Accord — Accomplir tâches efficacement	100 %	67 %	83 %
Accord — Autonomie professionnelle préservée	100 %	67 %	83 %
Accord — Apport concret dans la pratique	**100 %**	**17 %**	**67 %**
Accord — Compatible avec la fonction judiciaire	100 %	50 %	100 %
Accord — À l'aise de continuer à utiliser	100 %	83 %	83 %
Durée d'accès perçue comme suffisante	**71 %**	**17 %**	**83 %**
Score NPS moyen (/10)	**10,0**	**7,3**	**9,8**
Expérience positive ou très positive	**100 %**	**17 %**	**83 %**
Recommande la poursuite (Oui)	71 %	50 %	83 %
Recommande un élargissement quelconque du projet	86 %	50 %	100 %

5.7.2 Effet du retrait de l'accès (Groupe B)

Le Groupe B a répondu à deux questions spécifiques portant sur l'effet du retrait de l'accès avant la fin du projet.

Énoncé (Groupe B uniquement)	Accord
Le retrait de l'accès avant la fin du projet m'a permis de mieux mesurer	33 % (2/6)

Énoncé (Groupe B uniquement)	Accord
la valeur réelle de l'outil dans mon travail.	
Après le retrait de l'accès, j'ai constaté un manque dans certaines tâches que j'effectuais auparavant avec l'aide des agents IA.	33 % (2/6)

Ces résultats suggèrent que l'appropriation des outils était encore partielle au moment du retrait pour la majorité des membres du Groupe B. Seuls deux participants sur six ont constaté un manque après le retrait et estimé que cela leur avait permis de mieux en mesurer la valeur. La majorité est demeurée neutre, ce qui corrobore l'hypothèse que le retrait est intervenu avant que ces juges aient pu atteindre un niveau d'appropriation suffisant pour percevoir l'outil comme irremplaçable.

5.7.3 Effet de l'accès tardif (Groupe C)

Le Groupe C a répondu à une question portant sur l'effet de l'accès plus tardif sur la formation de leurs habitudes d'usage.

Énoncé (Groupe C uniquement)	Accord
Le fait d'avoir reçu l'accès plus tardivement a limité ma capacité à développer des habitudes d'usage significatives.	**0 % (0/6)**

Résultat contre-intuitif : la continuité prime sur le moment d'accès

Le résultat le plus saillant de cette analyse est que le Groupe C (accès tardif, mais ininterrompu jusqu'à la fin) obtient des résultats quasi équivalents au Groupe A (accès complet dès le début) et nettement supérieurs au Groupe B (accès retiré avant la fin) : NPS 9,8 vs 10,0 vs 7,3 ; 83 % d'expérience positive vs 100 % vs 17 %; 100 % recommandent un élargissement vs 86 % vs 50 %. Aucun membre du Groupe C (0 %) ne juge que l'accès tardif a limité sa capacité à développer des habitudes d'usage significatives. Ce résultat suggère que la continuité de l'accès est un facteur déterminant de l'adoption et de la satisfaction, et ce indépendamment du moment auquel cet accès commence. Pour les déploiements futurs, cette donnée plaide en faveur d'un accès simultané et ininterrompu pour l'ensemble des participants plutôt que d'une approche par phases de retrait.

5.8 Consommation de crédits Copilot

Sur l'ensemble de la durée du projet, incluant la phase de test préalable au déploiement officiel, les agents déployés ont généré une consommation cumulative de crédits Copilot répartie comme suit :

Agent	Crédits consommés
Assistant rédacteur	1 253
Traducteur	498

Agent	Crédits consommés
Code de procédure civile IA	247
Livre bleu 2.0 (droit de la famille)	128
Technicien informatique IA	84
Code civil du Québec IA	101
Citateur juridique	77
Code criminel IA	45
Loi sur la faillite et l’insolvabilité IA	23
Total	**2456**

Ces chiffres couvrent l’ensemble de la période du projet, y compris la phase de test préalable au déploiement officiel du 8 décembre 2025. La consommation est dominée par l’Assistant rédacteur (1 253 crédits, soit 51 % du total), suivi du Traducteur (498 crédits, 20 %) et du Code de procédure civile IA (247 crédits, 10 %), ce qui est cohérent avec les données d’utilisation présentées à la section 5.1.

Ces données constituent une référence utile pour estimer les besoins en crédits d’un déploiement élargi. Le groupe test comptant 22 juges, un déploiement à l’ensemble de la Cour supérieure pourrait potentiellement multiplier cette consommation par un facteur d’environ 10, selon les volumes d’utilisation observés. Cette projection devra être intégrée dans la planification budgétaire de toute extension du programme.

6. DISCUSSION

6.1 Efficacité et valeur ajoutée des agents IA

Les données recueillies indiquent que les agents IA ont généré une valeur ajoutée réelle et mesurable dans le contexte du travail judiciaire préparatoire. Quatre dimensions de valeur se dégagent de l'analyse des verbatims et des données quantitatives :

- Efficacité rédactionnelle : La reformulation, la révision et la traduction de textes constituent les usages dominants, avec un niveau de satisfaction élevé. Plusieurs juges ont mentionné que la reformulation proposée par l'IA surpassait parfois la version originale.
- Environnement sécurisé : La possibilité d'utiliser des outils d'IA dans un cadre institutionnellement sécurisé est perçue comme un avantage décisif, permettant d'éviter le recours à des outils commerciaux non contrôlés.
- Accélération de la recherche : Même si la fiabilité de la recherche juridique nécessite vérification, plusieurs participants ont souligné que l'IA accélère considérablement le survol des autorités et l'exploration des pistes de recherche.
- Décharge cognitive partielle : Le fait que 14 % des répondants à l'évaluation Teams rapportent une diminution de leur charge cognitive pour certaines tâches préparatoires est un indicateur positif, même si 65 % estiment que la charge est demeurée stable.

6.2 Limites et risques identifiés

Les participants ont identifié plusieurs limites et risques importants à prendre en compte pour la suite :

- Hallucinations : Le risque d'hallucination, soit la génération par un agent IA d'une information inexacte ou inexistante présentée comme étant valide, constitue la principale limite identifiée dans le cadre du projet pilote. Bien que ce phénomène soit inhérent à l'utilisation de modèles d'intelligence artificielle générative, son occurrence est demeurée très limitée dans le cadre de l'expérimentation.

  Moins d'une dizaine de cas ont été documentés sur l'ensemble des interactions et des tests réalisés, incluant les questions synthétiques. Rapporté au volume total d'utilisation (893 sessions), ce phénomène apparaît marginal et circonscrit. Les cas observés présentent par ailleurs des caractéristiques communes permettant d'en identifier les causes principales.

  L'analyse des occurrences indique que ces situations sont majoritairement attribuables à des facteurs contrôlables, notamment (1) des limites ou imperfections dans les bases de connaissances intégrées aux agents, (2) certaines contraintes ou ambiguïtés dans les instructions système, et (3) des requêtes complexes ou atypiques situées en périphérie du périmètre prévu pour l'agent. Ces constats suggèrent que les hallucinations observées relèvent moins d'un comportement aléatoire du modèle que de conditions spécifiques liées à son environnement de déploiement.
- Dépendance potentielle : Certains participants ont exprimé la crainte de développer une dépendance excessive aux outils, pouvant réduire la vigilance dans la vérification des informations.
- Concurrence d'outils plus performants : L'existence d'outils commerciaux plus puissants mais non contrôlés pourrait mener à un désintéressement des agents institutionnels si leur qualité n'est pas améliorée.
- Absence d'interconnexion entre agents : La nécessité de passer d'un agent à l'autre selon la tâche est perçue comme une friction importante par plusieurs participants.

- Paramétrage restrictif : Certains participants estiment que les paramètres restreints des agents ont considérablement limité l'usage, notamment pour la recherche juridique et la rédaction.

6.3 Analyse du phasage expérimental

L'analyse des données relative au phasage révèle des résultats nuancés et contre-intuitifs. Parmi les répondants du groupe B (accès retiré avant la fin), seuls deux participants sur six ont constaté un manque après le retrait, ce qui suggère que l'appropriation était encore partielle au moment du retrait pour la majorité d'entre eux. Pour le groupe C (accès tardif), aucun participant ne considère que le moment d'accès tardif ait limité leur capacité à développer des habitudes d'usage significatives, ce qui confirme que la continuité de l'accès importe davantage que son moment de début. Ces résultats sont analysés en détail à la section 5.7.

6.4 Indépendance judiciaire et IA

L'une des préoccupations centrales mentionnées dans les sondages de départ était l'impact potentiel de l'IA sur l'indépendance judiciaire. Les résultats du sondage final sont à cet égard rassurants : 84 % des répondants affirment que les agents leur ont été utiles sans nuire à leur autonomie professionnelle et à leur jugement. Aucun répondant ne rapporte d'expérience où l'IA aurait influencé une décision judiciaire.

6.5 Contexte médiatique et devoir de réserve

Au cours de la période entourant le projet pilote, les médias ont rapporté l'existence d'un jugement rendu en Cour supérieure qui ferait l'objet de préoccupations quant à l'utilisation potentielle de l'intelligence artificielle et au risque d'hallucinations. Ce dossier fait présentement l'objet d'un appel devant la Cour d'appel du Québec.

La Cour supérieure considère que son devoir de réserve l'empêche de commenter ce dossier ou de se prononcer sur le fond de l'affaire, qu'elle laisse entièrement à l'appréciation de la Cour d'appel. Cela dit, il ne lui est pas possible d'ignorer complètement cet élément dans le cadre du présent rapport, compte tenu de la résonance publique qu'il a suscitée et de sa pertinence apparente pour les questions que ce projet pilote cherche précisément à documenter.

> **Précision importante**
>
> Le jugement concerné a été rendu avant le début du projet pilote IA (déployé le 8 décembre 2025). Il ne saurait donc être attribué, de quelque manière que ce soit, aux outils ou aux pratiques encadrés dans le cadre du présent projet. Cette précision est essentielle pour éviter toute confusion entre l'usage libre et non encadré d'outils d'IA commerciaux grand public et le cadre institutionnel rigoureux mis en place par la Cour dans ce projet.

Les juges de la Cour supérieure ont par ailleurs été sensibilisés aux avantages et aux risques de l'intelligence artificielle lors d'une formation obligatoire tenue dans le cadre de l'Assemblée annuelle de la Cour supérieure en octobre 2024. Par ailleurs, l'utilisation de l'IA à l'extérieur de l'encadrement du projet pilote a été fortement déconseillée par la direction de la Cour. Le contexte médiatique entourant ce dossier illustre précisément pourquoi un encadrement institutionnel rigoureux est nécessaire : il ne s'agit pas d'empêcher les juges d'avoir accès aux outils d'IA, mais de s'assurer qu'ils y accèdent dans des conditions qui protègent l'institution, les justiciables et les juges eux-mêmes.

6.6 Enjeux institutionnels : l'absence de ressources techniques internes

Un enjeu structurel transversal au projet pilote mérite d'être explicité dans ce rapport : l'absence de ressources techniques internes spécialisées en intelligence artificielle ou en intelligence d'affaires au sein de la Cour supérieure. Cette lacune a complexifié chaque phase du projet (planification, déploiement, formation et soutien continu) et a constitué la principale contrainte opérationnelle lors du projet.

La Cour supérieure n'étant pas autonome au niveau de ses budgets et de sa structure de soutien, elle doit obligatoirement obtenir l'autorisation du ministère de la Justice (MJQ) avant de procéder à une telle embauche.

Il faut également souligner que le MJQ a apporté un soutien précieux tout au long du projet, notamment pour ce qui concerne l'environnement technologique Microsoft 365 et la validation sécuritaire. Cependant, la nécessité de protéger l'indépendance judiciaire a rendu difficile le partage libre d'information et le recours à un soutien de tiers extérieurs sur des questions touchant directement au fonctionnement des agents. Il importait en effet d'éviter que des choix techniques ou administratifs du MJQ viennent influer, même indirectement, sur les paramètres des agents ou sur les contenus auxquels les juges avaient accès : une telle influence aurait été contraire aux exigences de l'indépendance institutionnelle des tribunaux.

Cette expérience rend incontournable la nécessité, pour tout tribunal souhaitant déployer des agents IA de manière pérenne, de se doter d'une expertise technique interne dédiée. Cette expertise doit être suffisamment autonome pour assurer la conception, le paramétrage, les tests, la maintenance et l'évolution des outils sans dépendance structurelle vis-à-vis de partenaires gouvernementaux ou commerciaux extérieurs.

6.7 Les agents IA comme outils vivants : enjeux de maintenance continue

Contrairement à des ressources documentaires statiques, les agents IA déployés dans le cadre du projet pilote se sont révélés être des outils évolutifs, dont le fonctionnement dépend de paramètres techniques extérieurs qui peuvent changer indépendamment des choix opérationnels de l'institution. Cette nature « vivante » des agents a eu des conséquences concrètes sur la conduite du projet.

Ainsi, le déploiement prévu des agents a dû être retardé de plusieurs semaines au début de l'expérimentation lorsque le modèle de langage sous-jacent à Microsoft Copilot a été mis à jour, passant de GPT-4.0 à GPT-4.1. Cette mise à jour a modifié suffisamment le comportement des agents pour nécessiter de retester l'ensemble du parc d'agents avant tout déploiement, afin de s'assurer que les instructions système, les bases de connaissances et les restrictions configurées produisaient toujours les résultats attendus. Par la suite, des bogues survenus en cours d'expérimentation ont également obligé l'équipe à reparamétrer et retester certains agents, engendrant des interruptions de service temporaires et une charge de travail additionnelle significative pour l'équipe de coordination.

Implication pour les déploiements futurs

Le maintien opérationnel d'agents IA en contexte judiciaire ne peut être assimilé à la gestion d'une infrastructure statique. Il exige un suivi continu, une capacité de test rapide et une expertise suffisante pour intervenir dès qu'une mise à jour de la plateforme ou qu'un changement de comportement est détecté. Cette réalité opérationnelle doit être intégrée dès la phase de planification de tout projet similaire, avec des ressources humaines et des

processus de validation formellement attribués.

7. CONCLUSIONS ET RECOMMANDATIONS

7.1 Conclusions principales

Ce projet pilote a démontré la faisabilité et la pertinence de l'intégration d'agents IA dans le travail judiciaire préparatoire. Les principaux enseignements sont :

1. L'IA est viable dans un contexte judiciaire sous réserve d'un encadrement rigoureux. 84 % des participants affirment la compatibilité des agents IA avec les exigences de la fonction judiciaire.
2. Les usages les plus efficaces sont la révision/reformulation de texte, la traduction, et le soutien à la rédaction. La recherche juridique nécessite encore des améliorations de fiabilité.
3. L'environnement sécurisé est un différenciateur institutionnel clé qui justifie le maintien d'une solution propre à la Cour.
4. Le Compagnon CAIJ représente un outil complémentaire plutôt que concurrent, particulièrement performant pour la recherche juridique spécialisée.
5. Une majorité significative (89 %) des participants souhaite continuer à utiliser des agents IA dans un cadre balisé, et 63 % recommandent un déploiement élargi.

7.2 Recommandations

Recommandation 1 : Constitution d'une expertise technique interne

Cette recommandation constitue une condition de succès essentielle à la mise en œuvre de l'ensemble des recommandations qui suivent : sans ressource technique interne autonome, les autres orientations resteront difficiles à concrétiser de façon durable.

La pérennisation d'un programme d'agents IA au sein de la Cour supérieure nécessite le développement d'une expertise technique interne dédiée, indépendante des partenaires gouvernementaux ou commerciaux. Cette expertise devrait couvrir la conception et le paramétrage des agents, la constitution et la mise à jour des bases de connaissances, les protocoles de test et de validation, ainsi que le soutien aux utilisateurs. Cette autonomie technique est une condition nécessaire pour préserver l'indépendance judiciaire dans la gouvernance des outils IA. L'expérience du présent projet pilote démontre que l'absence d'une telle ressource constitue un risque opérationnel significatif, voire un empêchement à tout déploiement à plus grande échelle.

Il importe toutefois de rappeler que la Cour supérieure ne jouit pas d'une autonomie administrative complète : elle ne contrôle ni ses budgets ni ses processus d'embauche. La décision de créer un poste dédié à l'intelligence artificielle ou à l'intelligence d'affaires relève donc du ministère de la Justice du Québec (MJQ), qui doit en autoriser la création et le financement. Dans ce contexte, cette recommandation s'adresse autant à la direction de la Cour qu'au MJQ, dont la collaboration active est indispensable à sa mise en œuvre.

Recommandation 2 : Déploiement élargi

En regard des résultats obtenus, la Cour supérieure devrait envisager un déploiement élargi des agents IA, combiné à des produits complémentaires, dont le Compagnon CAIJ, à l'ensemble des juges volontaires de la Cour supérieure. Ce déploiement devrait être accompagné d'un programme de formation renforcé et d'un mécanisme d'évaluation continue.

Recommandation 3 : Amélioration des agents

Les agents devraient être améliorés sur trois axes principaux : (a) réduction des hallucinations dans les agents de recherche juridique, (b) interconnexion entre agents pour réduire la friction dans le passage d'un outil à l'autre, et (c) assouplissement des paramètres des agents de rédaction pour permettre des interactions plus complexes, tout en maintenant certaines balises (longueur de la boîte de contexte, instructions spécifiques, refus de prédire l'issue d'un litige, refus de produire une opinion juridique, etc.).

Recommandation 4 : Évaluation longitudinale

Un suivi longitudinal des indicateurs d'utilisation et de satisfaction devrait être mis en place pour mesurer l'évolution de l'adoption dans le temps et détecter tout effet pervers (dépendance, baisse de vigilance, etc.).

Recommandation 5 : Développement du cadre éthique

Le cadre de gouvernance devrait être mis à jour pour inclure des directives spécifiques sur les usages interdits, les procédures de vérification obligatoires, et les protocoles en cas d'erreur de l'IA identifiée. Ce cadre devrait être porté à la connaissance de l'ensemble des juges de la Cour.

Recommandation 6 : Partage inter-institutionnel

Les résultats de ce projet pilote devraient être partagés avec d'autres tribunaux provinciaux et fédéraux intéressés par des démarches similaires, afin de contribuer au développement d'un cadre commun pour l'utilisation de l'IA dans les institutions judiciaires canadiennes.

Recommandation 7 : Planification de la maintenance continue des agents

Tout déploiement futur d'agents IA doit intégrer, dès la phase de conception, un protocole de maintenance continue. Les agents IA sont des outils évolutifs dont le comportement peut être modifié par des mises à jour de la plateforme sous-jacente indépendamment de toute action de l'institution. Ce protocole devrait notamment prévoir : des cycles réguliers de tests après chaque mise à jour de la plateforme, un mécanisme de signalement rapide des anomalies par les utilisateurs, une procédure formalisée de suspension temporaire et de révision du paramétrage en cas de bogue, et des ressources humaines explicitement assignées à ces tâches. La maintenance des agents doit être considérée comme un coût récurrent permanent, et non comme une tâche ponctuelle de démarrage.

8. LIMITES DE L’ÉTUDE

Cette étude comporte plusieurs limites méthodologiques à considérer dans l’interprétation des résultats :

- Taille de l’échantillon : Avec 19 répondants au sondage final, la puissance statistique est limitée et les résultats ne peuvent être généralisés à l’ensemble des juges de la Cour supérieure.
- Biais de sélection : Les participants étaient volontaires, ce qui suggère un biais d’auto-sélection vers les juges les plus ouverts aux technologies.
- Durée du projet : La durée de 90 jours peut être insuffisante pour observer les effets à long terme de l’adoption ou les risques émergents.
- Absence de groupe de contrôle strict : Le design expérimental, bien que sophistiqué, ne permet pas une comparaison rigoureuse avec un groupe n’ayant eu accès à aucun outil IA.
- Métriques de plateforme : Les données d’utilisation collectées automatiquement par la plateforme Microsoft ne capturent pas la qualité des échanges individuels ni l’impact réel sur la qualité des décisions.

9. CONCLUSION

Le projet pilote d'intelligence artificielle constitue une avancée significative dans l'intégration réfléchie et encadrée de l'IA au sein de la Cour supérieure du Québec. Les résultats obtenus démontrent qu'il est possible de déployer des outils d'IA dans un environnement judiciaire de manière à la fois efficace et respectueuse des exigences fondamentales de la fonction : indépendance, impartialité, confidentialité et rigueur.

Les agents IA ont démontré leur utilité principalement pour les tâches de révision, reformulation, traduction et soutien à la rédaction. La majorité des participants souhaitent continuer à les utiliser et recommandent un déploiement élargi. Les limites identifiées, notamment le risque d'hallucinations et la nécessité de vérification systématique, sont bien comprises par les utilisateurs et gérées dans le cadre du projet.

Ce projet ouvre la voie à une transformation progressive, prudente et éclairée de la pratique judiciaire à l'ère de l'intelligence artificielle. Il fournit une base empirique solide pour les décisions institutionnelles à venir et un modèle de gouvernance reproductible pour d'autres institutions du système judiciaire canadien.

La Cour remercie chaleureusement les 22 juges qui ont accepté de participer à cette expérience, Me Guillaume Bourgeois, adjoint exécutif au bureau de la juge en chef, qui en a été l'idéateur et le coordonnateur, ainsi que la cellule CINTIA du ministère de la Justice du Québec pour son soutien technique et juridique tout au long du projet.

ANNEXES

Annexe A — Catalogue des agents déployés

Le projet pilote a mobilisé neuf agents IA, chacun configuré dans Microsoft Copilot Studio avec une base de connaissances spécifique, des instructions système dédiées et un format de sortie prescrit. Le tableau ci-dessous présente l'ensemble de ces agents.

Agent	Fonction principale	Base de connaissances	Format de sortie
Assistant rédacteur	Révision, reformulation, amélioration stylistique de textes préparatoires	Aucune (modèle linguistique général)	Texte révisé libre
Traducteur	Traduction français-anglais de textes juridiques et judiciaires	Aucune (modèle linguistique général)	Texte traduit intégral
Code civil du Québec	Recherche et localisation d'articles du CCQ	Texte intégral du Code civil du Québec	Reproduction verbatim des articles pertinents
Code de procédure civile	Recherche et localisation d'articles du CPC	Texte intégral du Code de procédure civile	Reproduction verbatim des articles pertinents
Code criminel	Recherche et localisation d'articles du Code criminel	Texte intégral du Code criminel (L.R.C. 1985)	Reproduction verbatim des articles pertinents
Loi sur la faillite et l'insolvabilité	Recherche d'articles de la LFI en droit commercial	Texte intégral de la LFI	Reproduction verbatim des articles pertinents
Citateur juridique	Correction formelle de citations selon la méthodologie interne	Guide interne du Service de recherche de la Cour	Citation réformée conformément au guide interne
Livre bleu 2.0 (droit de la famille)	Recherche dans l'ouvrage de doctrine interne en droit familial	Livre bleu de la Cour supérieure (droit de la famille)	Reproduction verbatim des sections pertinentes
Technicien informatique IA	Soutien technique, dépannage, analyse de captures d'écran	Documentation technique interne (environnement Teams / MJQ)	Liste structurée d'étapes de dépannage numérotées

Annexe B — Instruments de mesure

Quatre instruments de collecte de données ont été utilisés pour assurer la triangulation des résultats. Le tableau ci-dessous décrit chaque instrument, sa période d'administration et les principaux thèmes abordés.

Instrument	Modalité	N	Thèmes principaux
Sondage de départ (Groupes A et B)	Début du projet (déc. 2025)	n = 13	Connaissances préalables en IA ; expériences antérieures d'outils IA ; attentes ; préoccupations initiales (confidentialité, indépendance, charge de travail) ; niveau d'aisance technologique.
Sondage de départ (Groupe C)	Début du projet (déc. 2025)	n = 6	Connaissances préalables en IA ; expériences antérieures d'outils IA ; attentes ; préoccupations initiales (confidentialité, indépendance, charge de travail) ; niveau d'aisance technologique ; les questions prennent une dimension prospective.
Sondage à deux mois	Approx. deux mois après le début du projet (fév. 2026)	n = 17	Satisfaction globale ; gain de temps ; facilité d'intégration aux habitudes ; agents les plus utilisés ; comparaison avec le Compagnon CAIJ ; facteurs de satisfaction et d'insatisfaction ; charge cognitive.
Sondage final	Fin du projet (mars 2026)	n = 19	Échelle de Likert (14 items) sur l'utilité, la fiabilité, la compatibilité avec la fonction judiciaire et la gouvernance ; fréquence et contextes d'utilisation ; Net Promoter Score (NPS) ; intentions pour la suite (déploiement élargi) ; verbatims libres.
Questions synthétiques d'évaluation	Tout au long du projet par l'équipe de coordination	9 agents × 10 questions	Exactitude et citation verbatim des textes sources ; pertinence des réponses ; complétude de la couverture ; détection d'hallucinations factuelles ; qualité de la traduction (6 sous-critères) ; qualité de la reformulation (4 sous-critères).

Annexe C — Grilles d’évaluation par questions synthétiques

Trois grilles d’évaluation ont été employées selon la nature de l’agent évalué.

C.1 — Grille standard (agents à base documentaire)

Appliquée aux agents Code civil, Code de procédure civile, Code criminel, Loi sur la faillite et Livre bleu 2.0. Quatre critères par question :

Critère	Méthode d’évaluation	Pondération
Citation verbatim	Vérification de la correspondance exacte entre la réponse de l’agent et le texte source officiel	Oui / Non (critère binaire)
Pertinence élevée	Évaluation de la pertinence du passage cité par rapport à la question posée	Élevée / Partielle / Faible
Complétude	Mesure du degré de couverture de la réponse par rapport à l’ensemble des éléments attendus	Complète / Partielle / Incomplète
Score global /10	Évaluation qualitative intégrante des trois critères précédents, en tenant compte des nuances et de l’utilité opérationnelle de la réponse	0 à 10

C.2 — Grille spécialisée Traducteur (6 critères, n = 10 textes)

Fidélité au sens du texte source, registre et ton juridique, structure et ponctuation, adaptation culturelle et terminologique, fluidité et lisibilité, cohérence terminologique sur l’ensemble du texte. Chaque critère évalué sur 10.

C.3 — Grille spécialisée Assistant rédacteur (4 critères, n = 10 textes)

Correction grammaticale, fidélité au sens du texte original, qualité du vocabulaire juridique employé, qualité syntaxique et structurelle. Chaque critère évalué sur 10.

C.4 — Grille spécialisée Citateur (n = 10 questions posées, 10 évaluées)

Correction formelle de la citation selon le guide interne du Service de recherche de la Cour (incluant l’ordre des éléments bibliographiques, la ponctuation, les abréviations et le format des références neutres), présence de tous les éléments requis, absence de données fabriquées ou inventées. Score global 0 à 10, avec échec automatique (0/10) en cas de fabrication d’information factuelle (hallucination avérée).

Annexe D — Protocole de traitement des données et résidence de l'information

Le dispositif de protection des données mis en place dans le cadre du projet pilote reposait sur quatre piliers complémentaires.

Pilier	Description
Environnement sécurisé MJQ	Déploiement exclusif dans l'environnement Microsoft 365 du ministère de la Justice du Québec (MJQ), réservé aux juges, avec chiffrement de bout en bout et hébergement visant le territoire canadien. Environnement validé préalablement par des experts en cybersécurité qui conseillent les tribunaux québécois.
Modèle bac à sable	Les agents n'ont aucun accès automatique aux documents présents sur les appareils des utilisateurs. Chaque juge alimentait manuellement les agents en soumettant des extraits ou des questions. Les sessions et historiques de conversation ne sont pas partagés entre utilisateurs. L'environnement bac à sable n'est pas visible ni accessible à la direction de la Cour ni au personnel du MJQ : seul le juge utilisateur a accès à son propre historique de conversation.
Instructions aux participants	Les participants ont reçu des directives explicites leur demandant de ne pas soumettre aux agents des informations confidentielles relatives aux dossiers en cours, des renseignements personnels sur des justiciables, ni tout autre contenu couvert par le secret professionnel dans un contexte judiciaire.
Limitation des garanties	Malgré l'ensemble de ces mesures, il n'était pas possible de garantir à cent pour cent le traitement de l'information en sol canadien en raison de la nature distribuée des infrastructures infonuagiques. Cette limite a été communiquée explicitement aux participants. Le protocole était conforme aux orientations du MJQ en vigueur au moment du projet.