SUPERIOR COURT OF QUÉBEC

EVALUATION REPORT

Artificial Intelligence Pilot Project

Full Report • April 2026

Data Collection Period
December 2025 to March 2026

EXECUTIVE SUMMARY

From December 2025 to March 2026, the Superior Court of Québec carried out a pilot project to evaluate the use of artificial intelligence (AI) agents in judicial work. The project, conducted under the authority of the Office of the Chief Justice, is, to our knowledge, one of the first documented and controlled experiments in AI integration within a Canadian superior court.

Twenty-two (22) volunteer judges took part in the project, representing 11.17% of the Superior Court's bench. They were divided into three groups under a staggered-phase experimental protocol. Nine specialized AI agents were deployed through Microsoft Teams (Copilot Studio), covering writing assistance, translation, legislative research, legal citation, and technical support.

Key Findings

89% of respondents stated that the limitations of AI agents were understandable and manageable. 84% believe that AI is compatible with the requirements of the judicial function, provided there is a clear governance framework. 89% said they would be comfortable continuing to use this type of tool in a well-defined framework. 63% recommend a broader deployment. The most-used agent (writing assistant) accumulated 479 sessions with a 94% engagement rate.

The agents proved particularly effective for text revision and rewriting (89% of uses), translation (63%), writing support (63%), and legal research (53%). The limitations identified relate mainly to variable response quality in specialized legal research, the need for systematic verification, and the lack of interconnection between agents.

The report concludes that AI use is viable in a judicial environment, provided strict conditions of governance, training, and supervision are in place, and it sets out concrete recommendations for the next steps of the project.

1. INTRODUCTION AND CONTEXT

1.1 Institutional Context

The Superior Court of Québec is the province's court of original general jurisdiction. It has jurisdiction over civil, commercial, family, criminal, and public law matters. Superior Court judges carry out their duties in a demanding work environment, characterized by high caseloads, strict procedural requirements, and numerous preparatory tasks.

Against this backdrop, the rise of generative artificial intelligence technologies makes it incumbent on judicial institutions to explore what these tools offer rigorously and carefully. Courts in Canada and abroad have begun to take up this question, but few have run structured experiments with sitting judges.

1.2 Issues and Project Rationale

In the years leading up to the project, Superior Court judges expressed growing interest in digital tools that could make their preparatory work more efficient. Several were already using consumer AI tools (ChatGPT, Copilot, and others) on their own for various tasks. The absence of an institutional framework, however, posed significant risks for confidentiality, reliability, and judicial independence.

The pilot project therefore set out to provide a controlled framework for exploring these tools safely, while generating empirical data on their effectiveness and their compatibility with the requirements of the judicial function.

1.3 Pilot Project Objectives

The pilot project pursued four main objectives:

- Measure the actual use of AI agents by judges in their daily functions.
- Evaluate the satisfaction and perceived utility of the deployed agents.
- Identify the most relevant uses and limitations of the tools in a judicial context.
- Define a reproducible governance framework for potential broader deployment.

2. GOVERNANCE FRAMEWORK

2.1 Guiding Principles

The governance framework adopted by the Superior Court for this pilot project[1] is based on five fundamental principles:

- **Principle 1:** Judicial Independence

No AI agent may influence a judge's judgment or decision. Use is limited to preparatory and support tasks.

- **Principle 2:** Data Confidentiality and Security

The agents are deployed within the Court's secure Microsoft 365 infrastructure, ensuring that case data does not pass through uncontrolled external servers.

- **Principle 3:** Transparency and Verification

Judges are required to verify all information provided by the agents before using it in their work. The AI's response is considered a starting point, not a final result.

- **Principle 4:** Proportionality of Uses

The use of agents is limited to defined and approved categories of tasks, excluding in particular the final drafting of judgments.

- **Principle 5:** Reversibility and Continuous Evaluation

The project is designed to allow for discontinuation at any time, and its effects are continuously evaluated using robust methodological instruments.

2.2 Usage Framework

The AI agents deployed in the project are built with predefined parameters that limit their scope of operation. Each agent specializes in a specific domain and can only respond to queries related to its function. This modular architecture reduces the risk of drift or unintended uses.

Beyond the thematic scope of each agent, explicit instructions were built into the system parameters to constrain not only the subject matter handled but also the format of the expected outputs. Each agent was configured to produce only one defined type of response, matched to a specific operational need in the judicial environment. This formatting constraint is designed to ensure the reliability, reproducibility, and direct usefulness of the outputs. Only two agents, the Writing Assistant and the Translator, received broader instructions, since writing and translation needs cannot easily be confined to a single format.

This "output specialization" approach is a concrete algorithmic governance measure: it reduces the risk of interpretive hallucination, strengthens source traceability, and aligns with the principle of judicial independence by fully preserving the judge's own judgment.

A training session was held for participants at the start of the project, followed by recurring weekly meetings in a hybrid format (in-person and Microsoft Teams) to follow the experience and gather participant feedback.

2.3 Confidentiality Protection

[1]The AI governance framework of the Superior Court of Québec is available at the following address: https://coursuperieureduquebec.ca/fileadmin/cour-superieure/A_propos/2025-09_Cadre_gouvernance_IA_Cour_superieure_Quebec.pdf

One of the core concerns identified at the outset of the project was safeguarding the confidentiality of judicial records. The baseline survey showed that data confidentiality and security were among the most frequently cited concerns of participants. The chosen solution, based on Microsoft 365 and Copilot Studio, is part of a protocol established by the Ministry of Justice of Québec (MJQ) for the hosting and processing of judicial data.

A further architectural decision was important here: rather than building a tool that automatically draws from documents on users' devices, the agents were designed on a "sandbox" model that each judge had to populate manually. Access to each agent was reserved to the individual judge, with no sharing of sessions or conversation histories between users.

Despite these measures, it was not possible to guarantee that information would be processed entirely on Canadian soil. Judges were therefore expressly instructed not to submit to the agents any document or text containing personal information or confidential information relating to ongoing cases.

Key Advantage

Several participants specifically singled out the secure environment as the main advantage of the project's agents over consumer commercial AI tools, allowing them to use similar functionality within an institutionally reliable framework.

3. METHODOLOGY

3.1 Experimental Design

The pilot project officially launched on December 8, 2025, and ended on March 16, 2026, for a total experimental period of 98 days (roughly 14 weeks). This period includes the agent access phase, data collection, and administration of the final survey.

The project used a staggered-groups experimental design, inspired by "stepped-wedge" protocols common in health and social science research. This design supports within-group comparisons (before and after tool access) while limiting selection effects. It also lets us observe how withdrawing access affects the tool's perceived value.

Three groups were formed:

Group	Access Modality	N (final survey)
Group A	Full access from the start, throughout	7 (37 %)
Group B	Access withdrawn before project end	6 (32 %)
Group C	Access starting later in the project	6 (32 %)

3.2 Participant Selection and Profile

Participation in the project was voluntary. Forty-two (42) Superior Court judges expressed interest. Of these, 22 were selected to form the pilot group through a two-step procedure: candidates were first sorted into sub-groups defined by representativeness criteria (diversity of subject areas such as civil, criminal, family, commercial, and mixed; geographic distribution across Montréal, Laval, Saint-Maurice, and other districts; prior level of technological comfort), and then a random selection was made within each sub-group. The aim was to build a sample as representative as possible of the entire Superior Court bench, while limiting the biases associated with voluntary self-selection.

A control group was initially planned but ultimately not set up. The intensity of support required (weekly meetings, individual follow-up, and ongoing facilitation) made it unrealistic to run a control group in parallel with the same rigour. The experimental design was therefore adapted to rely on before-and-after comparisons specific to each participant.

Final survey respondents (n = 19) had the following profile:

- Subject areas: Mixed/multiple areas (47%), Civil (26%), Family (11%), Criminal (11%), Other (5%)
- Prior technological comfort: Average (42%), Fairly high (32%), Very high (21%), Fairly low (5%)
- AI use before the project (baseline survey, n=13): Yes occasionally (38%), Yes regularly (31%), Rarely (23%), Never (8%)

3.3 Measurement Instruments

Five data-collection instruments were used, allowing the results to be triangulated:

- **Instrument 1:** Baseline survey for the late-access group (n=6, December 2025): administered before the project began to the sub-group of participants who joined later, this survey measured baseline technological comfort, prior AI use, initial concerns, and expectations. The results show that 100% of respondents identified confidentiality as the main concern and 83% identified data security; 5 out of 6 were comfortable with Microsoft Teams, and 4 out of 6 had prior AI experience.
- **Instrument 2:** Main baseline survey (n=13, December 2025): measured prior knowledge, expectations, levels of technological comfort, and initial concerns.
- **Instrument 3:** Two-month evaluation survey (n=17, February 2026): measured early adoption, usage patterns, and satisfaction after an initial period of use.
- **Instrument 4:** Usage data: quantitative metrics collected automatically by the platform (sessions, engagement, response quality, user reactions).
- **Instrument 5:** Final Survey (n=19, March 2026): overall evaluation of the experience, measurement of perceived impact, and recommendations for the next steps. It was administered at the very end of the project.

The final survey comprised 42 questions, including 14 items on a 5-point Likert scale, 8 multiple-choice questions, a recommendation scale from 0 to 10 (adapted Net Promoter Score (NPS)), and open-ended questions gathering participants' written comments.

3.4 Agent Quality Assessment: Synthetic Questions

In addition to the surveys, a structured evaluation of the agents' response quality was carried out using synthetic questions prepared internally. A total of 90 evaluations were run across 9 specialized agents (10 questions per agent), using a methodology tailored to the nature of each agent.

For the document-based legal agents (Civil Code, Code of Civil Procedure, Criminal Code, Bankruptcy Act), each evaluation used three criteria: (1) verbatim citation of the expected article (yes/no); (2) relevance (low, medium, high); and (3) completeness of the response (incomplete, partial, complete), with a composite score out of 10. The Translator agent was evaluated against six specialized criteria (fidelity, fluency, register, terminological consistency, structure, and cultural adaptation). For the Writing Assistant, the criteria were grammar, syntax, legal vocabulary, and fidelity to meaning. The Citation Agent underwent a partial evaluation focused on citation correction per the Court's internal Research Services guide. For every agent, a follow-up prompt was allowed when the initial response was incomplete.

3.5 Weekly Support and Emerging Use Cases

Throughout the pilot project, participants received structured support on a weekly basis. Regular semi-structured meetings brought the participating judges together to ensure collective progress in adopting the tools.

3.5.1 Meeting Format and Objectives

Each weekly meeting served three complementary goals: (1) an awareness component covering the benefits and risks of generative AI, tailored to the specific challenges of the judicial context (hallucinations, bias, independence, confidentiality); (2) a practical demonstration component focused on concrete use cases identified through user testing; and (3) an open discussion in which judges could share their experiences, challenges, and discoveries.

This semi-structured format fostered peer learning: use cases emerging from one judge's practice were quickly shared and tested by the whole group. This rapid feedback loop between real-world use and training was an important factor in enriching the project.

3.5.2 Emerging Use Cases

Collective exploration over successive weeks surfaced and validated use cases that had not been anticipated in the original project design. The most significant are listed below:

Use Case	Description and Added Value
Recognition of Digitized Handwritten Notes	Agents were used to transcribe and structure handwritten notes that had been digitized (scanned), considerably reducing manual transcription time for judges.
Transformation of Text into PowerPoint Presentations	The ability to automatically convert narrative text into structured slides was identified as useful for preparing institutional and educational presentations.
Creation of Summary Tables with Integrated Passages and References	Agents were used to generate synthetic tables incorporating excerpts from legal texts with their references, facilitating source comparison or synthesis of authorities.
Screenshot-Based Troubleshooting (IT Technician Agent)	The IT Technician AI agent was used by submitting screenshots of technical problems directly to it, allowing it to visually analyze the situation and propose solutions without IT department intervention.

These emerging use cases illustrate the importance of structured support in any AI integration project: users gradually discover possibilities that go beyond the cases originally planned, provided they have a safe space in which to experiment and share. The regular cadence of meetings also made it possible to adjust the configuration of some agents mid-project to better match observed needs.

> **Methodological Contribution**
> The weekly support component is, to our knowledge, a distinctive feature of this pilot project compared with other AI experiments in the public sector. It combined rigorous evaluation with an organizational-learning approach, while ensuring that tool use developed within an ethical, institutionally defined framework.

4. DEPLOYED AI AGENTS

4.1 General Architecture

The AI agents were deployed through Microsoft Copilot Studio integrated with Microsoft Teams. This architecture has several advantages: native integration with the collaboration tool judges already use, data isolation within the Québec government's Microsoft 365 environment, and the ability to deploy specialized agents with embedded knowledge (legal documents, codes, and the like).

4.2 Agent Descriptions

Agent	Main Function
Writing Assistant	Writing assistance, reformulation, and stylistic improvement of legal texts and other preparatory documents.
Translator	French-English and English-French translation of legal passages, with attention to terminological equivalences.
Civil Code of Québec AI	Research and explanation of articles of the Civil Code of Québec, with citation of relevant provisions.
Code of Civil Procedure AI	Research and explanation of articles of the Code of Civil Procedure, with identification of applicable rules.
Criminal Code AI	Research in the Criminal Code, identification of applicable offences and penalties.
Bankruptcy Act AI	Specialized research in the Bankruptcy and Insolvency Act.
The Citation Agent	Correction and restructuring of legal source references.
IT Technician AI	Technical support for issues related to the use of the Court's digital tools.
Blue Book 2.0	Verbatim reproduction of relevant sections of this internal family law doctrine work, prepared for the exclusive use of judges. The agent performed no synthesis or reinterpretation of the content.

4.3 Agent Configuration

Agent configuration is central to the pilot project's reliability and security. Each agent was configured in Microsoft Copilot Studio under a common architecture, with customizations specific to each use case.

4.3.1 Technological Environment and Data Residency

The agents were deployed in the Microsoft 365 environment of the Ministry of Justice of Québec (MJQ), reserved for judges. That environment provides: (1) end-to-end encryption of all communications; (2) data hosting within Canadian territory; (3) agent isolation under a 'sandbox' model, without automatic access to documents on users' devices; and (4) native integration with Microsoft Teams, already used by judges. The environment was validated prior to launch by the cybersecurity experts who advise Québec courts.

4.3.2 Knowledge Bases and System Instructions

Each agent was provided with a specific knowledge base: the legal agents incorporate the full legislative texts (Civil Code of Québec, Code of Civil Procedure, Criminal Code, Bankruptcy and Insolvency Act), the Superior Court's Blue Book (family law), and internal documents prepared by judges and support staff. These documents were specially formatted to optimize processing by the AI, including Microsoft Word macros that make footnotes easier for the model to recognize by folding them into the running text. Detailed system instructions ("system prompts") were drafted for each

agent to shape the response style, require verbatim citation of sources, define the scope of admissible topics, and set out the expected behaviour when requests fall outside that scope.

This integration is carried out solely for internal purposes, to support the location of sources and legal research, and falls squarely within the exercise of judicial functions.

In this setting, the use of legislative texts is not reproduction for distribution or commercial purposes; it is a functional use inherent to the Court's mission.

The Court continues to rely on the official versions of the statutes, and the tools developed are not a substitute for them.

Web browsing was also disabled for every agent, and access to their default knowledge base (the general knowledge embedded in the underlying model) was blocked. The agents were thus directed exclusively to the designated knowledge bases, ensuring that their responses rely only on sources verified and validated by the coordination team.

4.3.3 Built-in Restrictions and Safeguards

In keeping with the governance framework, restrictions were built directly into the configuration of each agent. Generation capacity and context window length were deliberately limited to reduce the risk of hallucinations. The agents are configured to automatically decline requests for: assistance with the judicial decision-making process, generation of draft judgments, interpretation of legal texts, summarization of lengthy texts, and the drafting of legal opinions. These safeguards ensure that the agents remain support tools rather than substitutes for judicial reasoning. Each session opens with a reminder that all generated content must be verified.

The team is aware of the inherent limitations of system instructions: they add a further layer of protection but are not foolproof against determined users. They form part of a defence-in-depth approach that also relies on participant training, risk awareness, and the professional obligations specific to the judicial function.

5. RESULTS

5.1 Usage Data

Data automatically collected by the Microsoft Copilot Studio platform over the course of the project offer a quantitative picture of agent adoption. The table below sets out key indicators by agent. Note that certain satisfaction indicators could not be measured for some agents because of low user feedback rates.

Agent	Sessions	Growth	Engagement	Satisfaction	Key Observation
Writing Assistant	479	+565%	94%	76.9%	Most-used agent. High demand.
Translator	154	+267%	94%	89.6%	286 questions asked. Strong growth throughout the project.
Code of Civil Procedure	67	+49%	96%	N/A	78 questions. Stable and steady growth.
Citation Agent	33	+74%	79%	90.9%	Markedly increasing use over the period.
IT Technician AI	40	+25%	85%	90.9%	Increasing use for troubleshooting and screen captures.
Civil Code of Québec	54	+116%	85%	N/A	Usage doubled. 65 questions asked. Synthetic score: 9.0/10.
Blue Book 2.0	38	+36%	87%	N/A	41% of respondents used it (two-month survey). Perfect verbatim citation.
Criminal Code	17	−15%	41%	N/A	Slight decline. Complex questions. Score 7.8/10 with high variability.
Bankruptcy Act	11	−45%	45%	N/A	Very specialized use. Synthetic score: 9.0/10.

In all, 893 sessions were recorded on the platform during the 90-day evaluation period, across all agents. The Writing Assistant leads by a wide margin with 479 sessions, nearly three times the second-placed agent (the Translator, 154 sessions). That lead reflects judges' interest in tools for rewriting, revision, and stylistic improvement. The Translator posted the strongest growth (+267%), evidence of gradual, sustained adoption over the course of the project. The legal research agents (Civil Code, Code of Civil Procedure, Criminal Code, Bankruptcy Act) were also adopted, though at more modest volumes consistent with their specialized nature. Every agent shows positive growth, which points to cumulative adoption rather than a novelty effect.

5.2 Survey Results

5.2.1 Frequency and Usage Contexts

The final survey (n=19) shows that the vast majority of participants used the AI agents in their judicial practice, often quite frequently:

- Several times per week: 63% (12/19)
- Almost every working day: 16% (3/19)
- About once per week: 16% (3/19)
- Less than once per week: 5% (1/19)

These results show strong adoption for a first-generation pilot project, with 79% of participants using the agents at least several times a week.

The most frequent usage contexts are, in decreasing order:

- Text revision or reformulation: 89% (17/19)
- Translation or language support: 63% (12/19)
- Writing support: 63% (12/19)
- Legal or doctrinal research: 53% (10/19)
- Legislative or regulatory research: 42% (8/19)
- Summarization of texts or documents: 32% (6/19)

These data confirm that perceived value is highest for language and drafting tasks, and more moderate for specialized legal research.

5.2.2 Likert Scale Results (Final Survey)

The following table presents the results of the 14 Likert scale items from the final survey (n=19):

Statement	Agree*	Neutral	Disagree
AI agents helped me accomplish certain tasks more efficiently	**84%**	5%	11%
AI agents allowed me to explore research avenues more quickly	**47%**	26%	26%
AI agents contributed to improving the quality of my preparatory outputs	**58%**	26%	16%
AI agents proved useful for my preparatory and repetitive tasks	**47%**	37%	16%
AI agents integrated well into my daily work routine	**68%**	21%	11%
AI agents were useful without undermining my professional autonomy or judgment	**84%**	11%	5%
AI agents represented a concrete contribution to my practice	**63%**	26%	11%
AI agent responses are sufficiently reliable for preparatory use	**74%**	16%	11%
Significant verification is required before using AI agent responses in my work	**47%**	32%	21%
AI agents sometimes directed me toward incorrect or misleading avenues	**21%**	37%	42%
The limitations of AI agents are understandable and	**89%**	5%	5%

Statement	Agree*	Neutral	Disagree
manageable in my work context			
The usage framework established for the project was adequate	**74%**	0%	26%
AI agents are compatible with the requirements of the judicial function, provided a clear governance framework is in place	**84%**	16%	0%
I would be comfortable continuing to use this type of tool within a well-defined framework	**89%**	0%	11%

** Agree = "Somewhat agree" + "Strongly agree". Disagree = "Somewhat disagree" + "Strongly disagree".*

5.2.3 Recommendation and Overall Perception

The final survey also gathered participants' intentions and recommendations regarding the next steps of the project:

- 63% (12/19) recommend a broader deployment of the tool to all judges
- 16% (3/19) recommend a gradual expansion to other volunteer judges
- 16% (3/19) are uncertain about the next steps
- 5% (1/19) recommend extending the small-scale experimentation

> **Key Indicator**
> Combining the 63% who favour a broader deployment with the 16% who favour a gradual expansion, 79% of participants (15 out of 19) want to see AI use in their judicial practice continue and expand. That figure is a strong signal in favour of sustaining the programme, provided the agents continue to improve and the framework remains rigorous.

On the overall continuation recommendation, 68% (13/19) of respondents recommend it without reservation, 26% (5/19) are uncertain, and 5% (1/19) do not recommend it. This distribution, combined with the 79% rate in favour of some form of expansion, suggests that even the more hesitant participants recognize the project's value and envisage continuing it, though they prefer a more gradual approach.

5.3 Two-Month Evaluation (n = 17)

The two-month evaluation survey captured participants' experience after an initial period of using the agents in Teams. The main findings are as follows:

- Overall satisfaction: 47% satisfied, 29% moderately satisfied, 18% slightly satisfied, 6% very satisfied
- Time savings: 29% significant, 47% moderate, 12% little or very little
- Integration into work habits: 35% very easily, 29% easily, 24% moderately, 12% with difficulty
- Wish to continue after the project: 76% "Definitely yes", 12% "Probably", 12% "Maybe"
- Most-used agents: Writing Assistant (100%), Translator (71%), AI Laws (53%), Blue Book (41%)

The main drivers of satisfaction are ease of use (65% of respondents), speed (53%), and response quality (47%). The main limiting factors are variable response quality (47%) and lack of legal precision (47%).

5.4 Comparison with the CAIJ Companion

As part of the two-month survey, participants were asked to compare the pilot project's AI agents with the CAIJ Companion, an AI tool developed by the Centre d'accès à l'information juridique (CAIJ), to which they had beta access.

- Reliability of results for legal research: 71% consider the CAIJ superior, 18% consider them comparable
- Speed and efficiency: 41% CAIJ superior, 29% comparable, 24% AI agents superior
- Ease of integration into habits: 35% AI agents superior, 35% comparable
- Preference for legal research: 71% CAIJ, 18% both depending on context

Interpretation

These results suggest that the two tools are complementary rather than in competition, particularly on speed and integration with Teams. The CAIJ Companion is seen as more specialized for legal research, while the pilot project agents are valued for their ease of integration with Teams and their versatility (writing, translation, technical support).

5.5 Synthetic Question Results by Agent

The synthetic-question evaluation covered all 9 deployed agents. Two evaluation grids were used, depending on the agent: a standard grid (verbatim citation, relevance, completeness, score out of 10) for agents referencing source texts, and specialized grids for the Translator, Writing Assistant, and Citation Agent.

5.5.1 Document-Based Agents (n = 60 questions)

The six agents that reference legal or doctrinal texts directly were each put through 10 questions, evaluated on three criteria (verbatim citation, relevance, completeness) and an overall score out of 10.

Agent	Questions	Avg. Score	Verbatim Citation	High relevance	Completeness
Civil Code of Québec	10	**9.0 / 10**	100%	90%	70%
Bankruptcy Act	10	**9.0 / 10**	100%	80%	80%
Code of Civil Procedure	10	**8.9 / 10**	100%	70%	70%
IT Technician	10	**8.6 / 10**	N/A	90%	70%
Blue Book 2.0 (family law)	10	**7.0 / 10**	100%	90%	40%
Criminal Code	10	**7.8 / 10**	100%	80%	50%

The Civil Code of Québec and Bankruptcy Act agents earned the highest overall scores (9.0/10), with perfect verbatim citation. The Bankruptcy Act also shows the highest completeness rate among the document-based agents (80%). The Code of Civil Procedure (8.9/10) maintains 100% verbatim citation but a completeness rate of 70%, reflecting occasionally partial responses to complex procedural questions. The IT Technician (8.6/10), which does not cite source texts, shows 90% high relevance and 70% completeness. The Blue Book 2.0 agent (7.0/10) has perfect verbatim citation (100%) and high relevance in 90% of cases, but its completeness rate (40%) is the lowest in the group, indicating that responses correctly address core family law concepts but frequently lack a fully developed analytical structure. The Criminal Code agent (7.8/10) shows the greatest variability (range 3/10 to 9/10) and a completeness rate of 50%, reflecting the frequent complexity and ambiguity of criminal law questions.

5.5.2 Specialized Agents with Custom Evaluation Grids

The Translator agent (n = 10 texts) was evaluated according to six specialized criteria:

Evaluation Criterion (Translator)	Average Score /10	Min. Score
Fidelity to meaning	**10.0 / 10**	10.0
Register and tone	**10.0 / 10**	10.0
Structure and punctuation	**9.9 / 10**	9.0
Cultural adaptation	**9.8 / 10**	9.0
Fluency	**9.6 / 10**	9.0
Terminological consistency	**9.5 / 10**	8.0

The Translator agent delivered excellent results: fidelity to meaning and legal register scored a perfect 10/10 on every evaluated text. Terminological consistency is the most variable criterion (minimum 8.0/10), reflecting the complexity of equivalences between French and English legal terms.

The Writing Assistant agent (n = 10 texts) was evaluated on four criteria: Grammar (9.6/10), Fidelity to meaning (9.8/10), Legal vocabulary (9.0/10), Syntax (8.9/10). The strongest performance is on texts with simple stylistic requirements. Texts with syntactic ellipses or specialized technical terms produce slightly more variable results (minimum 7.0/10 for legal vocabulary).

The Citation Agent (n = 10 questions evaluated) averaged 8.8/10, with high relevance at 100% and completeness at 90%. The agent is very strong at formal citation correction according to the Court's internal Research Services guide, producing perfect results (10/10) in 9 out of 10 cases. The only notable exception involved a question about a journal article: the agent invented the author's first name, a factual hallucination that led to a score of 0/10. This isolated incident is telling: the agent is highly reliable for the kinds of citations it knows well, but it can fail unexpectedly on less-standardised doctrinal sources. Beyond that limitation, the results confirm the agent's reliability within its current scope while underscoring that its focus on purely formal correction limits its practical usefulness. The agent would benefit from being expanded with a bank of sources frequently cited before the Superior Court, which would in time enable content verification and retrieval of complete references.

Indicator	Result
Questions evaluated	10 questions (out of 11 asked)

Indicator	Result
Average score	8.8 / 10
Perfect scores (10/10)	9 questions out of 10
High relevance	100%
Full completeness	90%
Factual hallucination detected	**1 case: author's first name fabricated (score: 0/10)**

5.5.3 Comparative Summary and Cross-Sectional Analysis

The comprehensive evaluation of all 9 agents yields a clear performance ranking. The writing and translation agents lead the way: the Translator (9.8/10) and the Writing Assistant (9.3/10) deliver the most consistent and highest results. The Citation Agent (8.8/10) ranks just below, with excellent performance on standard citations but some vulnerability to factual hallucinations for less-indexed sources. The document-based legal agents fall in the 8.6 to 9.0/10 range, except for the Criminal Code (7.8/10), whose marked variability (3/10 to 9/10) reflects the complexity inherent in criminal law. The Blue Book 2.0 (7.0/10) is a particularly instructive case: despite perfect verbatim citation, its completeness rate (40%) is the lowest overall, suggesting that the agent faithfully reproduces relevant passages but does not always cover the full scope of the question, a shortfall in analytical structure rather than content. These findings align with the participants' written comments, which identify writing and translation as the most satisfying uses and specialised legal research as the area most in need of improvement.

> **Methodological Note**
>
> The synthetic questions are idealized scenarios: they were formulated to test specific competencies, and they were evaluated internally according to the subjective expectations of their author. In practice, judges ask more complex, more contextual, and often multidimensional questions, which may produce real-world performance below the scores seen in this structured setting. That said, the agents were consistently able to produce the verbatim text of the applicable provisions; it was on completeness and relevance that points were generally lost.

5.6 Supplementary Analysis: Synthetic Performance / Perceived Satisfaction Matrix

Comparing the synthetic evaluation scores (Appendix C) against perceived satisfaction indicators drawn from platform data and surveys lets us place each agent along two independent axes: objectively measured performance and user-perceived value. The analysis distinguishes agents whose perception is out of step with their measured performance (over- or under-rated) from those whose perception aligns with it.

Agent	Eval. Score /10	Platform satisfaction	Positioning	Analytical Observation
Translator	**9.8 / 10**	89.6%	*High perf., high satisfaction*	Optimal alignment: perceived value is consistent with measured excellence.
Writing Assistant	**9.3 / 10**	76.9%	*High perf., moderate satisfaction*	Under-rated by users. Despite the best composite score and the highest usage volume (479 sessions), platform satisfaction is the lowest, which may reflect higher expectations or more complex uses.
Civil Code of Québec	**9.0 / 10**	N/A	*High perf., satisfaction not measured*	Excellent synthetic score (verbatim citation 100%, relevance 90%). No per-agent satisfaction data on the platform.
Bankruptcy Act	**9.0 / 10**	N/A	*High perf., satisfaction not measured*	Score comparable to the CCQ with the best completeness rate in its group (80%).
Code of Civil Procedure	**8.9 / 10**	N/A	*High perf., satisfaction not measured*	Excellent synthetic score. No direct per-agent satisfaction data on the platform.
Citation Agent	**8.8 / 10**	90.9%	*High perf., high satisfaction*	Positive alignment. High satisfaction despite the detected hallucination, suggesting that users value overall reliability and may not notice edge cases.
IT Technician AI	**8.6 / 10**	90.9%	*Adequate perf., high satisfaction*	Slightly over-rated by users. Satisfaction is high relative to the evaluated score, likely because of the practical, immediate nature of the technical support provided.
Criminal Code	**7.8 / 10**	N/A	*Moderate perf., satisfaction not measured*	Marked variability (range 3/10 to 9/10), consistent with the complexity of

Agent	Eval. Score /10	Platform satisfaction	Positioning	Analytical Observation
				criminal law questions.
Blue Book 2.0	**7.0 / 10**	N/A	*Moderate perf., satisfaction not measured*	Lowest score in the group. Despite perfect verbatim citation, the completeness rate (40%) limits practical value. Use is nonetheless notable (41% at the two-month survey).

** Platform satisfaction: recommendation rate (thumbs up) from Microsoft Copilot Studio data. Engagement: rate of sessions that generated a user response. N/A: indicator not available at this level of granularity.*

Methodological Limitation: Feedback Fatigue

Satisfaction rates from the Microsoft Copilot Studio platform are based on voluntary user reactions (thumbs up/down). Feedback fatigue may have set in over time, with the most active participants cutting back on manual ratings for routine sessions. These figures should therefore be read as trend indicators rather than absolute measures of satisfaction.

Interpretation: The Writing Assistant Case

The Writing Assistant is the most instructive analytical case: it has the second-highest composite evaluation score, behind the Translator (9.3/10), by far the highest usage volume (479 sessions), yet the lowest platform satisfaction (76.9%). Two complementary hypotheses explain this gap between measured performance and perceived value: (1) users hold higher expectations for a generalist, heavily used agent, and (2) it is applied to more complex, context-dependent tasks, where the model's limits surface more readily than on structured tasks like translation or technical support. This pattern tracks the literature on AI tool adoption: the most-used tools are also the ones onto which users project the widest range of expectations.

5.7 Comparative Analysis by Experimental Group

The staggered-group design lets us compare participants' experiences by access arrangement: full access throughout (Group A, n = 7), access withdrawn before project end (Group B, n = 6), and access beginning later in the project (Group C, n = 6). The analysis draws on responses to the final survey (n = 19), with participants identifying their group through a dedicated survey question. The 3 non-respondents are not assigned to a specific group.

5.7.1 Key Indicators by Group

Indicator	Group A (Full Access)	Group B (Access Withdrawn)	Group C (Late Access)
N (final survey)	7	6	6
High frequency (≥ several times/week)	86%	67%	83%
Agree: accomplish tasks efficiently	100%	67%	83%
Agree: professional autonomy preserved	100%	67%	83%
Agree: concrete contribution in practice	**100%**	**17%**	**67%**
Agree: compatible with judicial function	100%	50%	100%
Agree: comfortable continuing to use	100%	83%	83%
Access duration perceived as sufficient	**71%**	**17%**	**83%**
Average NPS score (/10)	**10.0**	**7.3**	**9.8**
Positive or very positive experience	**100%**	**17%**	**83%**
Recommends continuation (Yes)	71%	50%	83%
Recommends any form of project expansion	86%	50%	100%

5.7.2 Effect of Access Withdrawal (Group B)

Group B answered two specific questions on the effect of having access withdrawn before the end of the project.

Statement (Group B Only)	Agree
Having access withdrawn before the end of the project helped me better gauge the real value of the tool in my work.	33% (2/6)
After access was withdrawn, I noticed a gap in certain tasks that I previously performed with the help of AI agents.	33% (2/6)

These results suggest that adoption of the tool was still partial at the time of withdrawal for most Group B members. Only two of the six participants noticed a gap after withdrawal and felt it helped them better gauge the tool's value. The majority remained neutral, supporting the hypothesis that withdrawal came before these judges had reached a level of adoption at which the tool felt irreplaceable.

5.7.3 Effect of Late Access (Group C)

Group C answered one question on the effect of later access on forming usage habits.

Statement (Group C Only)	Agree
Receiving access later limited my ability to develop meaningful usage habits.	**0% (0/6)**

Counter-Intuitive Finding: Continuity Matters More Than Timing

The most striking result of this analysis is that Group C (late access, but uninterrupted through the end) achieves results nearly on par with Group A (full access from the start) and well above Group B (access withdrawn before the end): NPS 9.8 vs 10.0 vs 7.3; positive experience 83% vs 100% vs 17%; recommend expansion 100% vs 86% vs 50%. No member of Group C (0%) felt that late access limited their ability to build meaningful usage habits. The takeaway is that continuity of access is the decisive factor for adoption and satisfaction, regardless of when that access begins. For future deployments, these data support giving all participants simultaneous and uninterrupted access rather than a phased-withdrawal approach.

5.8 Copilot Credit Consumption

Over the full duration of the project, including the pre-deployment testing phase, the agents accumulated the following Copilot credit consumption:

Agent	Credits Consumed
Writing Assistant	1,253
Translator	498
Code of Civil Procedure AI	247
Blue Book 2.0 (family law)	128
IT Technician AI	84
Civil Code of Québec AI	101
Citation Agent	77

Agent	Credits Consumed
Criminal Code AI	45
Bankruptcy and Insolvency Act AI	23
Total	**2,456**

These figures cover the full project period, including the testing phase prior to the official deployment on December 8, 2025. Consumption is driven by the Writing Assistant (1,253 credits, or 51% of the total), followed by the Translator (498 credits, 20%) and the Code of Civil Procedure AI (247 credits, 10%), which is consistent with the usage data presented in Section 5.1.

These figures provide a useful benchmark for estimating credit needs for a broader deployment. With a test group of 22 judges, a deployment to the entire Superior Court could increase consumption roughly tenfold, based on observed usage volumes. That projection will need to be built into the budget planning for any expansion of the programme.

6. DISCUSSION

6.1 Effectiveness and Added Value of AI Agents

The data collected show that the AI agents generated real, measurable added value in preparatory judicial work. Four dimensions of value emerge from the analysis of the open-ended comments and quantitative data:

- Writing efficiency: Rewriting, revision, and translation are the dominant uses, with high satisfaction levels. Several judges noted that the AI's proposed rewording sometimes improved on the original.
- Secure environment: Being able to use AI tools within an institutionally secure framework is seen as a decisive advantage, making it unnecessary to fall back on uncontrolled commercial tools.
- Faster research: Even though the reliability of legal research still requires verification, several participants noted that AI substantially speeds up the review of authorities and the exploration of research avenues.
- Partial cognitive offloading: That 14% of Teams-evaluation respondents report a drop in cognitive load for certain preparatory tasks is a positive signal, even though 65% feel that their workload remained stable.

6.2 Identified Limitations and Risks

Participants identified several important limitations and risks to consider going forward:

- Hallucinations: The risk of hallucination (an AI agent generating inaccurate or non-existent information and presenting it as valid) is the main limitation identified in the pilot project. Although this phenomenon is inherent to generative artificial intelligence models, it remained very limited during the experiment.

  Fewer than ten cases were documented across all interactions and tests, including synthetic questions. Relative to the total volume of use (893 sessions), the phenomenon appears marginal and contained. The observed cases also share common features that help identify the main causes.

  A case-by-case analysis indicates that these situations are mainly attributable to controllable factors: (1) limitations or gaps in the knowledge bases built into the agents; (2) constraints or ambiguities in the system instructions; and (3) complex or atypical requests at the edges of the agent's intended scope. These findings suggest that the observed hallucinations are less a matter of random model behaviour than of specific conditions in the deployment environment.
- Potential over-reliance: Some participants raised concerns about becoming overly reliant on the tools, which could reduce their vigilance in verifying information.
- Competition from more capable tools: More powerful but uncontrolled commercial tools could draw users away from the institutional agents if the quality of the latter is not improved.
- Lack of interconnection between agents: Having to switch from one agent to another depending on the task is seen by several participants as significant friction.
- Restrictive configuration: Some participants feel that the agents' tight parameters significantly limited use, particularly for legal research and writing.

6.3 Analysis of the Experimental Phasing

The data on phasing yield nuanced and counter-intuitive findings. Among Group B respondents (access withdrawn before the end), only two of six participants noticed a gap after withdrawal, suggesting that adoption was still partial at the time of withdrawal for most. For Group C (late access), no participant considers that the timing of late access limited their ability to develop meaningful usage habits, confirming that continuity of access matters more than when it begins. These results are analyzed in detail in Section 5.7.

6.4 Judicial Independence and AI

One of the central concerns mentioned in the baseline surveys was the potential impact of AI on judicial independence. The final survey results are reassuring in this regard: 84% of respondents state that the agents were useful to them without harming their professional autonomy and judgment. No respondent reports an experience where AI influenced a judicial decision.

6.5 Media Context and Duty of Reserve

During the period surrounding the pilot project, media reported on a Superior Court judgment that raised concerns about the potential use of artificial intelligence and the risk of hallucinations. This matter is currently before the Québec Court of Appeal.

The Superior Court considers that its duty of reserve prevents it from commenting on that case or ruling on the merits, which it leaves entirely to the Québec Court of Appeal. At the same time, the matter cannot be wholly set aside in this report, given the public attention it has drawn and its apparent relevance to the very questions this pilot project set out to document.

> **Important Clarification**
>
> The judgment in question was rendered before the start of the AI pilot project (deployed on December 8, 2025). It cannot therefore be attributed, in any way, to the tools or practices governed by the present project. This clarification is essential to avoid any confusion between the free and unregulated use of consumer commercial AI tools and the rigorous institutional framework put in place by the Court in this project.

Superior Court judges also received awareness training on the benefits and risks of artificial intelligence at a mandatory training session held during the Superior Court's Annual Assembly in October 2024. The use of AI outside the pilot project framework was also strongly discouraged by the Court's administration. The media context surrounding this matter illustrates precisely why rigorous institutional governance is necessary: the goal is not to prevent judges from accessing AI tools, but to ensure that they do so under conditions that protect the institution, the litigants, and the judges themselves.

6.6 Institutional Issues: The Absence of Internal Technical Resources

One structural challenge ran through the entire pilot project and deserves to be stated plainly: the Superior Court has no internal technical resources specialized in artificial intelligence or business intelligence. That gap complicated every phase of the project (planning, deployment, training, and ongoing support) and was the main operational constraint throughout.

Because the Superior Court does not have full autonomy over its budget or support structure, it must obtain authorization from the Ministry of Justice (MJQ) before making any such hire.

It should also be acknowledged that the MJQ provided valuable support throughout the project, particularly on the Microsoft 365 technological environment and security validation. The need to protect judicial independence, however, made it difficult to share information openly and seek support from external third parties on matters directly affecting how the agents operate. It was essential to prevent the MJQ's technical or administrative choices from influencing, even indirectly, the configuration of the agents or the content to which judges had access: such influence would have been inconsistent with the institutional independence of the courts.

This experience makes it clear that any court wishing to deploy AI agents on a permanent basis must build dedicated internal technical expertise. That expertise must be autonomous enough to handle the design, configuration, testing, maintenance, and evolution of the tools without structural dependence on external governmental or commercial partners.

6.7 AI Agents as Living Tools: Challenges of Continuous Maintenance

Unlike static documentary resources, the AI agents deployed in the pilot project turned out to be evolving tools whose behaviour depends on external technical parameters that can change independently of the institution's operational choices. This "living" nature of the agents had concrete consequences for how the project had to be run.

At one point, the planned deployment of the agents had to be pushed back by several weeks at the start of the experiment, when the language model underlying Microsoft Copilot was updated from GPT-4.0 to GPT-4.1. The update altered the agents' behaviour enough to require retesting the full set of agents before any deployment, to confirm that the system instructions, knowledge bases, and configured restrictions still produced the expected results. Later, bugs that surfaced during the experiment required the team to reconfigure and retest certain agents, which caused temporary service interruptions and significant added workload for the coordination team.

> **Implications for Future Deployments**
>
> The operational maintenance of AI agents in a judicial context cannot be equated with managing a static infrastructure. It requires continuous monitoring, rapid testing capacity, and sufficient expertise to intervene as soon as a platform update or behavioural change is detected. This operational reality must be integrated from the planning phase of any similar project, with human resources and validation processes formally assigned to it.

7. CONCLUSIONS AND RECOMMENDATIONS

7.1 Main Conclusions

This pilot project has shown that integrating AI agents into preparatory judicial work is both feasible and relevant. The main lessons are as follows:

1. AI is viable in a judicial context, provided rigorous governance is in place. 84% of participants confirm that AI agents are compatible with the requirements of the judicial function.
2. The most effective uses are text revision and rewriting, translation, and writing support. Legal research still needs improvements in reliability.
3. The secure environment is a key institutional differentiator and justifies maintaining a Court-specific solution.
4. The CAIJ Companion is a complementary rather than competing tool, and is particularly effective for specialized legal research.
5. A significant majority of participants (89%) wish to continue using AI agents within a defined framework, and 63% recommend a broader deployment.

7.2 Recommendations

Recommendation 1: Building Internal Technical Expertise

This recommendation is a key condition for the success of every recommendation that follows: without a dedicated internal technical resource, the other initiatives will be difficult to sustain over the long term.

Sustaining an AI agent programme within the Superior Court requires developing dedicated internal technical expertise, independent of governmental or commercial partners. That expertise should cover agent design and configuration, knowledge base creation and maintenance, testing and validation protocols, and user support. Such technical autonomy is a necessary condition for preserving judicial independence in the governance of AI tools. The experience of this pilot project shows that the absence of such a resource is a significant operational risk, and indeed an obstacle to any large-scale deployment.

The Superior Court, however, does not enjoy full administrative autonomy: it controls neither its budget nor its hiring. Creating a position dedicated to artificial intelligence or business intelligence therefore rests with the Ministry of Justice of Québec (MJQ), which must authorize and fund it. This recommendation is thus addressed as much to the Court's administration as to the MJQ, whose active collaboration is essential to carrying it out.

Recommendation 2: Broader Deployment

In light of the results, the Superior Court should consider a broader deployment of the AI agents (combined with complementary products, including the CAIJ Companion) to all volunteer Superior Court judges. That deployment should be paired with a strengthened training programme and an ongoing evaluation mechanism.

Recommendation 3: Agent Improvement

The agents should be improved on three main fronts: (a) reducing hallucinations in the legal research agents; (b) interconnecting the agents to reduce friction when switching between tools; and (c) loosening the parameters of the writing agents to allow for more complex interactions, while maintaining key guardrails (context window length, specific instructions, refusal to predict case outcomes, refusal to produce legal opinions, and so on).

Recommendation 4: Longitudinal Evaluation

A longitudinal follow-up of usage and satisfaction indicators should be put in place to track how adoption evolves over time and detect any adverse effects (over-reliance, reduced vigilance, and so on).

Recommendation 5: Development of the Ethical Framework

The governance framework should be updated to include specific directives on prohibited uses, mandatory verification procedures, and protocols for handling identified AI errors. The framework should be shared with all judges of the Court.

Recommendation 6: Inter-Institutional Knowledge Sharing

The results of this pilot project should be shared with other interested provincial and federal courts pursuing similar initiatives, to help develop a common framework for the use of AI in Canadian judicial institutions.

Recommendation 7: Planning for Ongoing Agent Maintenance

Any future deployment of AI agents must build in, from the design phase, a protocol for ongoing maintenance. AI agents are evolving tools whose behaviour can be changed by updates to the underlying platform, independently of any action by the institution. The protocol should include, among other things: regular testing cycles after each platform update; a fast-track mechanism for users to report anomalies; a formal procedure for temporary suspension and reconfiguration review in the event of a bug; and human resources explicitly assigned to these tasks. Agent maintenance must be treated as a permanent recurring cost, not a one-time start-up task.

8. STUDY LIMITATIONS

This study has several methodological limitations to keep in mind when interpreting the results:

- Sample size: With 19 respondents to the final survey, statistical power is limited and the results cannot be generalized to all Superior Court judges.
- Selection bias: Participants were volunteers, which suggests a self-selection bias toward judges most open to technology.
- Project duration: 90 days may be too short to observe the long-term effects of adoption or emerging risks.
- Absence of a strict control group: The experimental design, though sophisticated, does not allow a rigorous comparison with a group that had no access to any AI tools.
- Platform metrics: Usage data automatically collected by the Microsoft platform do not capture the quality of individual exchanges or their actual impact on decision quality.

9. CONCLUSION

The artificial intelligence pilot project is a significant step forward in the thoughtful and governed integration of AI within the Superior Court of Québec. The results show that AI tools can be deployed in a judicial environment in a way that is both effective and faithful to the core requirements of the judicial function: independence, impartiality, confidentiality, and rigour.

The AI agents proved their usefulness mainly for revision, rewriting, translation, and writing support. Most participants wish to continue using them and recommend a broader deployment. The limitations identified, notably the risk of hallucinations and the need for systematic verification, are well understood by users and managed within the project framework.

The project paves the way for a gradual, cautious, and informed transformation of judicial practice in the age of artificial intelligence. It provides a solid empirical basis for the institutional decisions to come and a reproducible governance model for other institutions within the Canadian judicial system.

The Court extends its warm thanks to the 22 judges who took part in this experience, to Me Guillaume Bourgeois, Executive Counsel to the Office of the Chief Justice, who initiated and coordinated the project, and to the CINTIA unit of the Ministry of Justice of Québec for its technical and legal support throughout.

APPENDICES

Appendix A: Catalogue of Deployed Agents

The pilot project deployed nine AI agents, each configured in Microsoft Copilot Studio with a specific knowledge base, dedicated system instructions, and a prescribed output format. The table below presents all of these agents.

Agent	Main Function	Knowledge Base	Output Format
Writing Assistant	Revision, reformulation, and stylistic improvement of preparatory texts	None (general language model)	Free revised text
Translator	French-English translation of legal and judicial texts	None (general language model)	Full translated text
Civil Code of Québec	Research and location of Civil Code of Québec articles	Full text of the Civil Code of Québec	Verbatim reproduction of relevant articles
Code of Civil Procedure	Research and location of Code of Civil Procedure articles	Full text of the Code of Civil Procedure	Verbatim reproduction of relevant articles
Criminal Code	Research and location of Criminal Code articles	Full text of the Criminal Code (R.S.C. 1985)	Verbatim reproduction of relevant articles
Bankruptcy and Insolvency Act	Research of Bankruptcy and Insolvency Act (BIA) articles in commercial law	Full text of the BIA	Verbatim reproduction of relevant articles
Citation Agent	Formal correction of citations according to internal methodology	Court's internal Research Services guide	Citation corrected in accordance with the internal guide
Blue Book 2.0 (family law)	Research within the internal family law doctrinal work	Superior Court Blue Book (family law)	Verbatim reproduction of relevant sections
IT Technician AI	Technical support, troubleshooting, screenshot analysis	Internal technical documentation (Teams/MJQ environment)	Structured numbered list of troubleshooting steps

Appendix B: Measurement Instruments

Four data collection instruments were used to ensure triangulation of results. The table below describes each instrument, its administration period, and the main themes covered.

Instrument	Modality	N	Main Themes
Baseline Survey (Groups A and B)	Project start (Dec. 2025)	n = 13	Prior AI knowledge; previous AI tool experience; expectations; initial concerns (confidentiality, independence, workload); level of technological comfort.
Baseline Survey (Group C)	Project start (Dec. 2025)	n = 6	Prior AI knowledge; previous AI tool experience; expectations; initial concerns (confidentiality, independence, workload); level of technological comfort; questions take a prospective dimension.
Two-Month Survey	Approx. two months after project start (Feb. 2026)	n = 17	Overall satisfaction; time savings; ease of integration into habits; most-used agents; comparison with CAIJ Companion; satisfaction and dissatisfaction factors; cognitive load.
Final Survey	Project end (March 2026)	n = 19	Likert scale (14 items) on utility, reliability, compatibility with the judicial function, and governance; frequency and usage contexts; Net Promoter Score (NPS); intentions for next steps (broader deployment); open-ended written comments.
Synthetic Evaluation Questions	Throughout the project by the coordination team	9 agents × 10 questions	Accuracy and verbatim citation of source texts; response relevance; coverage completeness; detection of factual hallucinations; translation quality (6 sub-criteria); reformulation quality (4 sub-criteria).

Appendix C: Evaluation Grids by Synthetic Questions

Three evaluation grids were used depending on the nature of the agent evaluated.

C.1: Standard Grid (document-based agents)

Applied to the Civil Code, Code of Civil Procedure, Criminal Code, Bankruptcy Act, and Blue Book 2.0 agents. Four criteria per question:

Criterion	Evaluation Method	Weighting
Verbatim Citation	Verification of the exact correspondence between the agent's response and the official source text	Yes / No (binary criterion)
High relevance	Evaluation of the relevance of the cited passage relative to the question asked	High / Partial / Low
Completeness	Measurement of the degree of response coverage relative to all expected elements	Complete / Partial / Incomplete
Overall Score /10	Integrative qualitative evaluation of the three preceding criteria, taking into account nuances and the operational utility of the response	0 to 10

C.2: Specialized Evaluation Grid for the Translator (6 criteria, n = 10 texts)

Fidelity to source text meaning, legal register and tone, structure and punctuation, cultural and terminological adaptation, fluency and readability, terminological consistency throughout the text. Each criterion evaluated out of 10.

C.3: Specialized Evaluation Grid for the Writing Assistant (4 criteria, n = 10 texts)

Grammatical correctness, fidelity to the original text's meaning, quality of legal vocabulary used, syntactic and structural quality. Each criterion evaluated out of 10.

C.4: Specialized Evaluation Grid for the Citation Agent (n = 10 questions asked, 10 evaluated)

Formal correction of the citation according to the Court's internal Research Services guide (including the order of bibliographic elements, punctuation, abbreviations, and the format of neutral references), presence of all required elements, absence of fabricated or invented data. Overall score from 0 to 10, with automatic failure (0/10) in the event of fabrication of factual information (confirmed hallucination).

Appendix D: Data Processing Protocol and Information Residency

The data protection system established for the pilot project rested on four complementary pillars.

Pillar	Description
MJQ Secure Environment	Exclusive deployment in the Microsoft 365 environment of the Ministry of Justice of Québec (MJQ), reserved for judges, with end-to-end encryption and hosting targeting Canadian territory. Environment validated prior to launch by cybersecurity experts who advise Québec courts.
Sandbox Model	Agents have no automatic access to documents on users' devices. Each judge manually populated the agents by submitting excerpts or questions. Sessions and conversation histories are not shared between users. The sandbox environment is not visible or accessible to the Court's administration or MJQ staff: only the judge-user has access to their own conversation history.
Instructions to Participants	Participants received explicit directives asking them not to submit to the agents confidential information relating to ongoing cases, personal information about litigants, or any other content covered by professional secrecy in a judicial context.
Limitation of Guarantees	Despite these measures, full processing of information on Canadian soil could not be guaranteed, given the distributed nature of cloud infrastructure. This limitation was clearly communicated to participants. The protocol complied with MJQ guidelines in effect at the time of the project.