

Generative AI: Opportunities to enhance CLEO's Guided Pathways

Progress report



Acknowledgments

This discussion paper is the result of a joint research project undertaken by CLEO (Community Legal Education Ontario/Éducation juridique communautaire Ontario) in partnership with McGill University and the Montreal Cyberjustice Laboratory.

The following CLEO researchers contributed to the research and writing:

Erik Bornmann (lead)
Samantha Hills
Joanne Lu
Elizabeth Robinson
Dylan Wall
Maya Willard-Stepan

The following McGill researchers contributed to the research and writing:

Prof. Fabien Gélinas (lead)
Mia Engelmann
Gillian Hunnisett
Nicole Pede
Anika Singh
Kendra Wong
Arthur Zimmerman Barbare

AI consulting and development for this project was done by Tangowork.

We are grateful to The Law Foundation of Ontario for funding this research project.

While financially supported by The Law Foundation of Ontario, Community Legal Education Ontario/Éducation juridique communautaire Ontario (CLEO) is solely responsible for all content.

Toronto: Community Legal Education Ontario, November 2025



Contents

Executive summary	4
Background	6
Access to justice context.....	6
The Generative AI opportunity	6
Project scope	7
Partnerships.....	8
Progress-to-date	9
Approach.....	9
Key activities.....	9
Emerging insights	13
Positive findings.....	13
Challenges.....	14
Summary	17
Priorities and next steps	18
Finalize functionality.....	18
Add privacy protections	18
Engage communities and end users	19
Launch user support chatbot	19
Operationalize and deploy	20
Scale across pathways.....	20
Incorporate file uploads	20
Evolve suggestion functionality into AI Coach.....	21
Conclusion: A careful balance of innovation and caution	22
Appendix A: Activity report	24
Appendix B: Scoring for benchmarking framework	28



Executive summary

This report provides an update on the progress of CLEO's Generative AI (GenAI) initiative, which aims to enhance its Guided Pathways tool. It covers developments since the release of the October 2024 discussion paper, outlining achievements, challenges, and next steps through August 2025.

The project is driven by CLEO's commitment to improving access to justice, particularly for people who are self-represented or have limited help with their legal problems. The Guided Pathways simplify complex legal processes by deconstructing court forms and other legal documents into question-and-answer-style interviews. However, the resulting document outputs can lack narrative flow. Effective storytelling is essential in legal advocacy, particularly in affidavits, pleadings, and other written submissions. Can GenAI safely and responsibly improve the clarity, persuasiveness, and completeness of these narratives while maintaining user control and trust?

To explore this question, CLEO developed the Narrative Assistant, a GenAI-powered prototype designed to help users craft clear, coherent, and persuasive legal narratives. The Narrative Assistant represents the first practical application under CLEO's broader GenAI initiative and offers early insights that will guide future development.

Over the first year, the project team has made substantial progress. The team has:

- refined GenAI prompts to improve narrative organization and persuasiveness.
- established a research and evaluation framework that includes a benchmarking process and structured scoring rubrics. The framework also integrates the Langfuse platform for tracking of AI tests and includes over twenty inter-rater reliability rounds to calibrate scoring and feedback.

- conducted over three hundred controlled tests and varied-input experiments for two CLEO Guided Pathways: family law emergency motions and Small Claims Court scenarios.
- developed and tested a three-part Suggestion Framework that includes a Claim Checker, a Story Checker, and an Evidence Checker. The framework prompts users to provide useful details in their narratives, without giving legal advice.
- introduced a Highlight Changes feature as a safeguard to identify and explain GenAI edits, making them more transparent to users.
- developed a preview version of a GenAI-powered chatbot that provides 24/7 user support within Guided Pathways to address common technical questions.

The findings to date are promising: the GenAI consistently improves clarity, coherence, and persuasiveness compared to baseline outputs. The project has effectively eliminated hallucinations and introduced meaningful safeguards to facilitate user review, such as the Suggestion Framework and Highlight Changes Feature. However, ensuring that the GenAI consistently retains all facts that are critical to the claim remains a challenge at times. Maintaining accuracy in how critical facts and their surrounding contextual details are represented is another area for improvement. Human evaluation has been vital for the accuracy and reliability of scoring GenAI outputs, while also driving iterative improvements in prompting and feature design. Overall, the insights confirm the value of the project's "test, evaluate, and iterate" approach.

The project will next focus on finalizing prompt sets and the Suggestion Framework for CLEO's family law emergency motion, Small Claims Court, and peace bond pathways. The project team will also strengthen the evaluation framework by refining the prompting of large language model (LLM) evaluation to improve the reliability of output issue detection and facilitate scalability.

The project will also begin evaluating privacy enhancing technology (PET) in the system while scaling the Narrative Assistant across a broader range of CLEO's Guided Pathways. This will involve systematically testing, evaluating, and refining narrative and suggestion prompts for each new pathway to ensure the quality and safety of outputs. The project will engage community partners and other stakeholders to gather feedback, ensure the tools reflect user needs, and build trust in their reliability and responsible design. Ultimately, CLEO aims to implement a vetted, scalable, and maintainable set of GenAI features within its Guided Pathways, which are supported by robust governance and privacy protections.



Background

Access to justice context

Across Ontario, thousands of people each year must represent themselves in legal matters, often with limited or no access to affordable legal services. For these individuals, court forms and legal procedures can be intimidating, time-consuming, and confusing.

CLEO's Guided Pathways address part of this challenge. They guide users step-by-step through online question-and-answer-style interviews, helping them fill out legal forms, understand their rights, and take their next steps. This structured interview approach, called decision trees or expert systems, makes completing complex legal documents more accessible.

Although the Guided Pathways excel at form completion, they are limited in their ability to help users tell their story. Many legal processes require more than filling in blanks in a form. They demand narratives that clearly and persuasively explain the facts. For people who self-represent, this storytelling task is often daunting and even prohibitive. Narratives generated through decision trees can be technically correct and thematically organized, but lack chronological flow and appropriate emphasis. This can detract from the coherence and persuasive force of the narrative. These limitations affect people's ability to present their case effectively.

The Generative AI opportunity

GenAI presents an opportunity to strengthen these narratives. Unlike rule-based document assembly used by CLEO's Guided Pathways, GenAI can synthesize information into continuous, human-like prose. Used responsibly, it could transform decision-tree outputs into clearer, more compelling legal narratives. This would allow litigants to present their facts more effectively without requiring professional drafting skills.

At the same time, CLEO acknowledges the risks associated with emerging technologies. GenAI systems can omit critical details, introduce inaccuracies, or generate instructions that sound like legal advice. They also raise privacy concerns related to the security of personal information. Any integration must be cautious, transparent, and guided by robust evaluation and privacy protections.

Can GenAI be safely and effectively integrated into CLEO’s Guided Pathways as an optional tool to improve narrative quality? And can this be done while maintaining user trust and protecting against harm?

Project scope

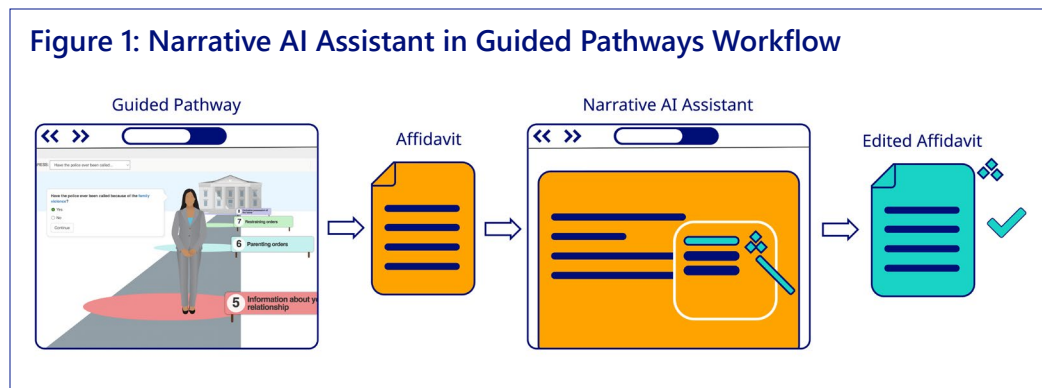
In October 2024, CLEO released a discussion paper outlining our GenAI initiative. We identified three possible uses for GenAI in the CLEO Guided Pathways:

1. **Narrative Assistant:** Improving the quality of narratives (i.e. affidavits, pleadings, supporting statements, etc.) generated from Guided Pathways. This is the current and primary focus.
2. **AI chatbot support:** Providing real-time technical assistance to users as they navigate the pathways.
3. **Data analysis tools:** Analyzing Guided Pathways usage data, identifying common user challenges, and improving service delivery.



While all three uses remain part of the project, the **Narrative Assistant** has been the focal point to date.

As shown in Figure 1, the Narrative Assistant takes the affidavit output by the Guided Pathways and generates an edited version. This edited affidavit organizes the user’s facts into a more logical flow and presents their arguments in a way that is more persuasive and coherent.



The Assistant also provides follow-up suggestions, by way of questions, to help the user further improve their affidavit.

The Narrative Assistant helps users tell their story persuasively, moving beyond simplifying task complexity to assist with drafting narrative.

While progress on the Narrative Assistant is the main subject of this report, the project team has also worked on the development of a new AI chatbot to support users. AI data analysis tools remain an area for future development.

Partnerships

CLEO has partnered with the McGill University Faculty of Law and the Montreal Cyberjustice Laboratory, with additional support from AI consultants at Tangowork. Together, these partners worked alongside CLEO's team of lawyers, analysts, and testers, as well as a McGill law professor and a team of law students, to form the full project team.

The project is presently in its first year (2025) of a three-year research and development initiative. It is supported by dedicated funding from the Law Foundation of Ontario which allows for iterative testing, consultation, and development.



Progress-to-date

Approach

The project began with exploratory testing of GenAI on the [family law emergency motion pathway](#). This pathway was chosen for two reasons:

- affidavit narrative quality can be critical in urgent family matters
- this pathway offered structured inputs suitable for AI experimentation

As work progressed, the team extended testing to [Small Claims Court claim pathways](#) and a [peace bond pathway](#).

Source material for testing involved synthetic data: hypothetical affidavits and pleadings developed to reflect existing CLEO Guided Pathway data.

CLEO also provided student testers with hypothetical scenarios, and these testers created affidavits and pleadings by using the Guided Pathways. Some testers intentionally replicated conditions in the source material that real users might experience. For example, they added spelling or grammatical errors to simulate users whose first language is not English. Sometimes testers left questions unanswered or marked “I don’t know” to mirror incomplete or uncertain responses. This approach ensured that the testing process captured not only controlled test conditions but also the variability and challenges of genuine user input.

The GenAI Narrative Assistant was first asked to draft an affidavit based on the input and then to evaluate its own drafts for clarity and completeness.

Key activities

Project work was structured in four phases. A detailed report of activities in each phase is included in Appendix A.

Initial setup and early testing (Oct-Dec 2024)

The project team conducted initial tests using a hypothetical emergency motion question-and-answer set. These early runs played with different models, model configuration settings (such as temperature and maximum output length), and variations in system instructions. This provided the team with a baseline understanding of how the GenAI behaviour shifted under different conditions.

Structured experimentation (Jan-April 2025)

Structured experiments continued using the same hypothetical emergency motion scenario as a stable baseline. The team then expanded the experiments to include new synthetic affidavits that were carefully drafted to reflect a cross section of deidentified outputs produced by pathways users. CLEO introduced a formal benchmarking framework with scoring across five categories:

- structure
- accuracy
- argumentative strength
- clarity
- omission of relevant details

The complete benchmarking framework is included in Appendix B.

The Narrative Assistant prototype integrated with Langfuse, an observability platform for AI systems. Langfuse automatically logs each test run, capturing the prompt, the AI's output, and the associated LLM evaluation. This integration was significant because it created transparency, enabled systematic evaluation, and supported scalability.

The project team also began exploring how GenAI could provide suggestions in the form of follow-up questions. After the Narrative Assistant generated an initial draft, testers were prompted to add or clarify details in their affidavits. For example, a suggestion might ask them to explain when a key event occurred or to expand on how a particular outcome affected them. The goal was not to change the draft itself, but to provide users with a structured way to edit and improve the GenAI-generated affidavit by adding more comprehensive information. Towards the end of this phase, the project team began to introduce legal reasoning into the suggestion prompting.

Expansion and iteration (May–June 2025)

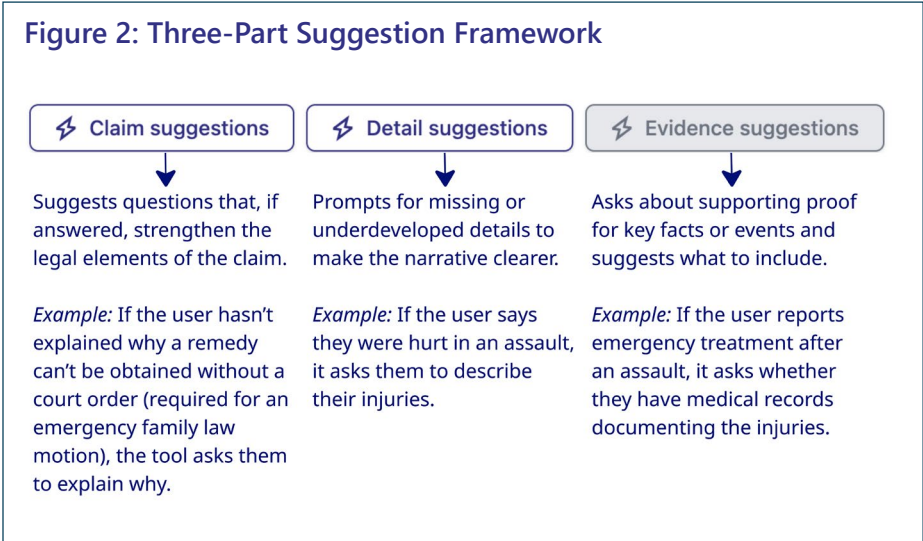
CLEO extended testing to include the Small Claims Court pathway that facilitates a claim for unpaid services. Weekly calibration sessions strived to improve inter-rater reliability of human scoring by reducing subjectivity. They also provided a forum for the team to refine and strengthen the evaluation rubric.

Building on results from earlier phases, the project team introduced new prompts for affidavit and claim drafting that embedded legal reasoning.

The suggestion system was formalized into a three-part Suggestion Framework focused on different aspects of the narrative:

- a Claim Checker which focuses on the elements of the claim or application
- a Story Checker which identifies problems or gaps in the story details
- an Evidence Checker which makes suggestions to include missing evidence.

Each part asked questions that, if answerable and answered, could strengthen the legal narrative outputs.



Evaluation and transparency enhancements (July–Aug 2025)

The project team enhanced the LLM evaluation with revised prompting to narrow the gap between AI and human scoring, and ensure consistent, systematic, and scalable evaluation.

The prototype site was updated to include a new Highlight Changes feature that automatically identifies relevant GenAI edits, additions, or omissions, and includes notes that explain each change. The feature shows users exactly how their text was modified and reduces the risk of discrepancies between the user's

intended meaning and the edited version of the affidavit. It provides greater transparency and avoids undermining the user's credibility.

Figure 3: Highlight Changes Feature

The screenshot shows a document titled "ORDERS REQUESTED" with several numbered items. Item 18 is highlighted in yellow and has a circled '1' next to it. Item 19 is also highlighted in yellow and has a circled '1'. Item 20 has a blue arrow pointing to a highlighted phrase and a circled '2' next to it. A pop-up box next to item 20 explains the change: "Changed order to 'stay at least 500 metres away from the children' from 'stay 500 metres metres away from the children.'" Below the pop-up are two buttons: "Ok, got it" and "Edit". Item 22 is highlighted in yellow and has a circled '1'. At the bottom, two callout boxes provide further details: Callout 1 states "Highlighted text identifies sections that have been modified by AI. Modifications include omitted text, added text as well as edited text." Callout 2 states "Pop ups explain each highlighted phrase that has been changed by AI. The user can choose to approve the change or make their own edits."

The project team compiled and reviewed the Claim Checker and Story Checker outputs to confirm their safety and relevance across multiple test runs. The Evidence Checker suggestions feature was left for development at a later time.

Work also began on a prototype chatbot to provide real-time support within the Guided Pathways. The team created a content management system, curated material from CLEO's public legal information resources, and launched an internal preview for testing common user technical support queries.

By the end of August 2025, the project had moved from initial exploratory tests to a structured, repeatable research program. Processes for benchmarking, prompt refinement, tester training, data logging, and user-facing safeguards were all established, laying the foundation for scaling the Narrative Assistant across additional pathways and deepening engagement with stakeholders.



Emerging insights

Throughout the first year of research and testing, several key insights have emerged. These insights provide a clearer picture of the potential of GenAI in CLEO's Guided Pathways, as well as the limits and conditions under which it can safely add value.

Positive findings

Improvements in narrative quality

Testing has consistently shown that GenAI improves the clarity, structure, and argumentative strength of legal narratives compared to benchmark Guided Pathway outputs created using decision trees. Affidavits and pleadings generated by the Narrative Assistant tend to:

- break complex sentences into simpler forms
- organize facts into a more logical flow
- present arguments in a way that appears more persuasive and coherent

For users of CLEO's Guided Pathways, these changes help them to present their story in a way that aligns with legal norms.

Structured inputs and context enhance performance

One of the most promising findings is that GenAI performs best when it is provided with carefully crafted decision tree outputs and guided by legal context. Detailed decision tree outputs, such as those created using A2J Author, reduce the noise of unanswered questions and irrelevant details. And when these decision tree outputs are processed with prompts enhanced with relevant legal reasoning, they produce narratives that are more accurate and persuasive.

For example, when provided with a framework about what courts consider urgent in emergency motions, the GenAI was able to emphasize the right details. This suggests that GenAI is not a substitute for legal expertise, but a tool that can amplify impact, clarity, and structure when provided with sound legal scaffolding.

Mostly safe and helpful suggestions

Testing shows that the GenAI generates follow-up questions that are safe and useful for strengthening the family law emergency motion affidavits. In two evaluation sets, the team reviewed about eighty suggestions in round one and roughly sixty suggestions in round two. While all suggestions were considered sufficiently safe, points were deducted from the few suggestions that raised concerns that the tone wasn't fully trauma-informed, a question felt slightly leading, or the relevance of the suggestion wasn't clear.

Even when relevance was imperfect, suggestions typically prompted constructive reflection and added detail. Analyzing the narrative output against a codified legal checklist, helps the GenAI flag missing facts or legal elements and present targeted prompts. This approach keeps analysis anchored in the user's story while using expert criteria, in the form of job aids, at scale. Current emphasis focuses on the use of a Claim Checker, which has shown to be the most effective method of generating high quality suggestions to date. Early results indicate this method can extend to the Story Checker and Evidence Checker, including additional inputs, such as uploaded documents or later updates, to further strengthen narratives.

Transparency builds trust

Stakeholder confidence depends on ensuring that users remain in control of their own story. The project continuously works to facilitate user autonomy through the design and function of features. For instance, the above-described Suggestion Framework operates so that the user makes the substantive change in their own words. The Highlight Changes feature is designed to show users how their text is modified by GenAI. It automatically marks edits, additions, and removals, with notes that explain each change. Testers evaluating the preliminary version of this new feature found it to be a promising safeguard that promotes user agency and trust. While improving omission detection remains an area for refinement, the feature shows potential as a helpful tool for ensuring clarity and accountability in GenAI-assisted drafting.

Challenges

While the project has made substantial progress in its first year, it has also confronted challenges inherent to both GenAI and access-to-justice innovation. Each challenge has required careful adjustments in approach, tools, or safeguards to ensure the project remains on track and aligned with CLEO's mission.

Preserving accuracy and completeness

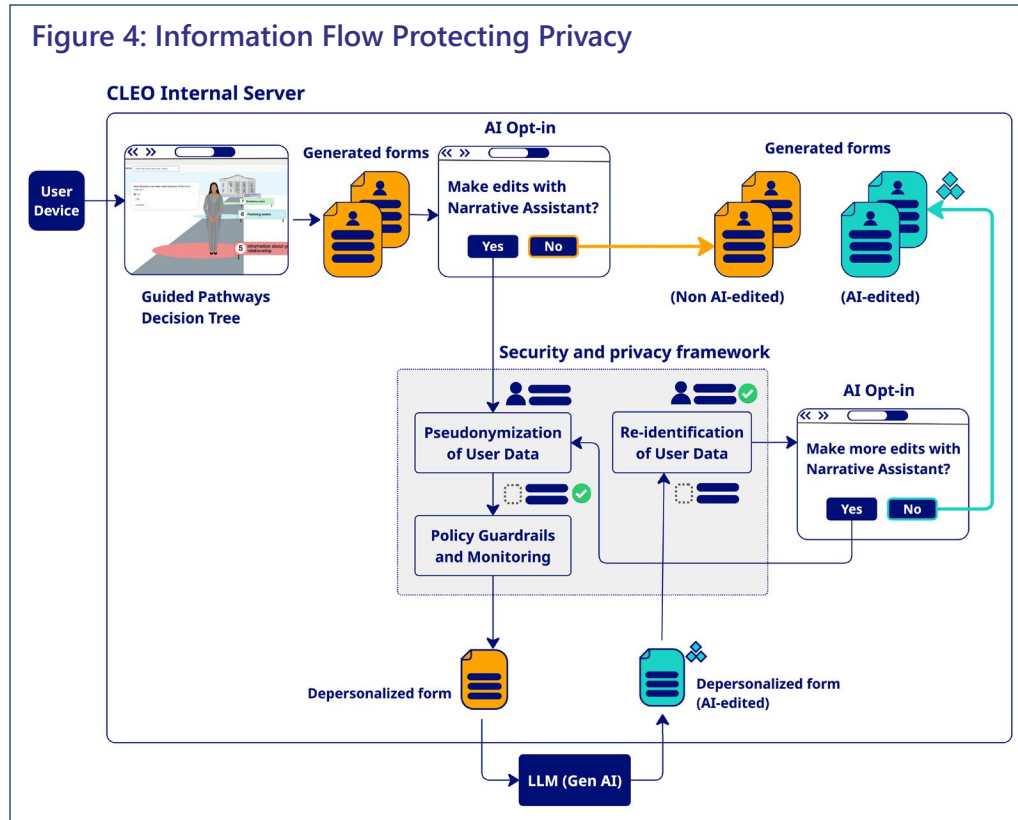
A consistent challenge has been ensuring that every critical fact provided by the user is accurately preserved in the AI-generated narrative. Although the GenAI produces more fluent and persuasive text, early and repeated tests revealed an occasional tendency to either omit important details or modify them in ways that created inaccuracies. These issues pose a direct risk to litigants' cases if essential information is lost or distorted.

The project team has responded by iteratively refining prompts to include legal reasoning and explicit instructions to include all relevant information and contextual details. This is complemented by the Highlight Changes feature that flags every GenAI edit, addition, or omission. The research team conceptualized and developed this safeguard to empower users to determine if essential facts had been altered or lost.

Protecting user privacy

CLEO's plan to use LLMs raises privacy questions that require further investigation. Narrative text can include indirectly identifying personal details, which are difficult to reliably recognize and remove. For example, even without names, narrative details like being 'the only English teacher in a small town' could still reveal the user's identity because the role itself is unique within that community. The project is exploring ways to ensure that no raw personal details leave CLEO, as shown in Figure 4.

Figure 4: Information Flow Protecting Privacy



Perfect anonymizing may not be possible with longer form narratives, so any remaining risk must be defined and checked regularly. The challenge is that removing too much detail can hurt usefulness or accuracy. And removing too little detail can risk users’ privacy. Users must have clear choices to opt in, opt out, and delete their data. Other safeguards include contracts prohibiting vendors from training on or keeping users’ data and storing data on CLEO’s Canadian servers. Finally, attacks that compromise the security of the system impose the need for stress tests and audits. CLEO must assess and determine that these protections are sufficient before releasing these GenAI tools for public use.

Variability of inputs and outputs

GenAI variability shows up on both sides: the same prompt can yield different responses and users supply wildly different inputs. These inputs can be highly detailed, very sparse, or occasionally incorrect. This unpredictability complicates evaluation and can erode confidence, especially when low-detail inputs lead the model to infer facts or add context that is not accurate. To assess this, the project adopted batch testing: multiple runs on the same input to confirm that, even when outputs differ, they remain safe, reliable, and legally sufficient. The project team iterated prompt configurations and model parameters to dampen harmful

fluctuations and discourage unsupported inferences, while preserving harmless stylistic variation. Regular human scoring meetings aligned evaluator judgment, helping testers consistently distinguish acceptable differences from risks affecting accuracy or completeness. Together, these practices turn variability from a source of uncertainty into a measured signal for quality control, improving robustness across the diverse ways people use Guided Pathways.

Scalability of evaluation

Scaling hinges on making AI evaluation of AI outputs reliable at production speed. The volume and variability of tests and live drafts quickly outstrip manual tracking and judgment, risking missed issues and inconsistent scoring. Integrating Langfuse to log prompts, outputs, and evaluator decisions allowed the project team to:

- centralize experimentation
- enable aggregated analysis
- create the telemetry needed for real-time monitoring across multiple pathways

Early “LLM as evaluator” trials, however, scored more generously than humans. This caused the project team to redefine rubric categories, iterate scoring prompts, and then reintroduce LLM evaluation alongside human review. Discrepancies have since narrowed, but full alignment remains a work in progress. Human oversight is still needed to ensure that users are provided with a safe and reliable tool.

The path to scale is phased. First, the project team plans to keep humans in the loop while calibration continues toward acceptable parity. Then, the LLM evaluation will be elevated to a primary role with targeted human quality assurance. This approach reduces reviewer burden and makes continuous, pathway-specific evaluation feasible at scale.

Summary

Overall, the first year of research demonstrates that GenAI has the potential to enhance the effectiveness of CLEO Guided Pathways outputs by making these narratives more coherent and compelling. It also confirms that responsible use requires legal context, rigorous evaluation frameworks, and transparency.

Challenges related to completeness, variability, persuasiveness, scale, evaluation, and privacy have necessitated systematic adjustments. Each adjustment has improved reliability and trustworthiness, reinforcing CLEO’s commitment to responsible innovation.



Priorities and next steps

Finalize functionality

The immediate priority is to finalize the core elements of the Narrative Assistant. This includes:

- strengthening narrative output generation
- refining user-facing suggestions
- improving the Highlight Changes feature
- bringing the LLM evaluation into closer alignment with human review

Completing this work will establish a stable base for implementing the Narrative Assistant and scaling its use to additional Guided Pathways.

Add privacy protections

As the project shifts closer to live use, protecting user information will be critical. Plans include implementing secure data handling and information governance frameworks, and considering privacy-enhancing technologies (PETs). These safeguards are prerequisites for responsible adoption.

Although our test data has been anonymized, any live Narrative Assistant would process real personal details. To protect privacy, we plan to explore a layered, built-in approach. This approach would automatically remove or mask personal details before AI processing. It would keep any personal identifiers separate and temporary and only reconnect them on CLEO's systems.

We will also explore strict data-use limits. This could include no public sharing of user data and contracts that prevent AI providers from training on or retaining user data. This would need to be accompanied by clear internal policies and procedures, such as user notices and choices (opt-in, withdrawal, deletion) and standard security practices. On the technical side, CLEO plans to evaluate combined methods to detect personal information, apply consistent pseudonymization, and strengthen quality checks. All these measures would be

carefully tested, including adversarial testing, to verify that privacy protections hold up to real use.

To stay current on fast-moving privacy technologies, CLEO has joined the Vector Institute. This is a Toronto-based, non-profit AI research institute focused on machine learning and deep learning, and part of Canada's Pan-Canadian AI Strategy. Through Vector's industry collaboration on detecting personal information, and other partnerships, CLEO will continue to explore privacy-enhancing technologies that could strengthen protections around image and document uploads.

Engage communities and end users

Meaningful engagement with the people who will use the Narrative Assistant will guide the next phase, with a consultation approach grounded in our partnerships with community organizations, subject-matter experts, and frontline legal service providers. CLEO recognizes that it will be challenging to reach some core audiences directly. Where direct input from users isn't feasible, we'll rely on trusted community partners to share perspectives and insight from the people they serve. Our goal is to gather practical, representative feedback that improves prompts, safeguards, and deployment choices so the tools are usable, safe, and trusted.

Early efforts will continue with organizations supporting women and children navigating family law after abuse, and with people using Small Claims Court Guided Pathways. This will take the form of short interviews, usability tests, and in-product feedback with users. As the AI expands to other pathways, we will broaden our consultation efforts.. We will be transparent about constraints, share back what we learn, and adjust plans based on partner feedback.

Launch user support chatbot

CLEO will launch an AI chatbot to serve users of the Guided Pathways with technical support and pathways-related public legal information. It will answer navigation and troubleshooting questions 24/7 using retrieval-augmented generation from a support knowledge base. It won't access pathway data unless the user shares it by entering it directly into the chatbot, and it will provide only technical help and pathway referrals, with clear disclaimers. Users will be cautioned to use a secure form connected to the pathways system to share any personally identifiable information with CLEO staff, not the chatbot.

Operationalize and deploy

The Narrative Assistant, with its Langfuse and anticipated PET components, will need to be rebuilt directly into the CLEO Guided Pathways system. It will also need to be deployed as part of a system update to the Guided Pathways to make it available to the public. This process will include upgrading the current pathways hosting system and conducting a privacy impact assessment and threat risk assessment (“PIATRA”) as a precondition for making these tools available for public use.

Scale across pathways

Once the foundation is secure, the project team will focus on expanding the Narrative Assistant across CLEO’s Guided Pathways where narrative detail most affects outcomes. Initial candidates include family law pathways where clear, fact-specific affidavits are essential. This might include:

- separation matters
- responding to or replying on applications
- motions (including urgent and change motions)
- trial-management steps

We also see value in adding the Narrative Assistant to criminal law pathways like private prosecutions and peace bonds, which require concise accounts tied to legal criteria. In rental housing, the assistant could help tenants articulate maintenance and repair issues or respond to eviction for unpaid rent. For Small Claims Court, it could support drafting and refining narratives in Plaintiff’s Claims and Defences. CLEO will continue to adapt prompts, suggestions, and safeguards to maintain accuracy and reliability as the system scales.

Incorporate file uploads

As we expand capabilities, a key next step is to explore document and image uploads for both the Narrative Assistant and the AI chatbot.

For the Narrative Assistant, uploads could extend the “make a suggestion” features: Claim Checker, Story Checker, and Evidence Checker. When the AI detects missing facts or weak support, it could invite the user to upload a related file, such as a letter, court document, photo, or receipt. It would then analyze the file and ask targeted follow-up questions to strengthen the narrative.

This research and development would proceed very cautiously and in anticipation of incorporating privacy-enhancing technologies. Any pilot would need to align with CLEO's privacy-by-design posture, including safeguards for removing or masking personal details, strict limits on data use and retention, and clear user consent and choice before the tool is made available to the public.

Evolve suggestion functionality into AI Coach

CLEO is interested in exploring a privacy-first, AI Coach supported with pathway inputs. This tool might help users with a range of tasks that could include:

- turning structured Guided Pathways inputs into clearer, more complete legal narratives
- surfacing targeted "make a suggestion" prompts drawn from curated expert checklists, combined with uploaded documents and new information
- preparing users for hearings and follow-through

The tool might take the form of an end-to-end coaching experience consistent with CLEO's principles of ease-of-use, reliability, inclusion, and robust privacy and security. A checklist-guided, suggestion-driven coach could help users navigate the "advice desert" between filing and hearing, strengthening their preparedness for the proceeding.



Conclusion: A careful balance of innovation and caution

The first phase of CLEO's GenAI project has demonstrated both the promise and the complexity of responsibly introducing advanced technologies into access-to-justice tools.

Since the October 2024 discussion paper, the project has moved through exploratory trials to structured experimentation, with broader applications and tangible innovations, including the Highlight Changes feature, improved Suggestion Framework, and a prototype chatbot. Each of these developments responds directly to the project's original vision: helping people tell their stories more clearly, while ensuring transparency, safety, and user control.

The research confirms that GenAI can enhance the quality of Guided Pathways outputs. Narratives produced with GenAI assistance are generally clearer, more persuasive, and easier to read than those generated solely through the Guided Pathways. At the same time, the project has identified some limits as well, including:

- risks of omission
- occasional accuracy problems
- difficulties in aligning AI evaluation with human judgment

By openly addressing these issues and embedding safeguards, such as iterative prompt refinement, transparency tools, and secure data handling, the project continues to strike a careful balance between innovation and caution.

Looking forward, the next stage will focus on consolidating the foundational components of the Narrative Assistant (i.e. narrative outputs, the Suggestion Framework, the Highlight Changes feature, and LLM evaluation) and prepare to scale it across many of CLEO's Guided Pathways. Community and stakeholder engagement will play a vital role in ensuring that the system is practical and

reliable. Strengthening privacy protections and expanding the chatbot will also be key priorities.

In conclusion, the project has progressed from a theoretical vision to a tangible prototype, establishing the groundwork for a responsible, reliable and scalable application of GenAI in public legal education. The coming phase will determine how these foundations translate into a sustainable tool that improves access to justice while maintaining the highest standards of transparency, reliability, and user trust.

Appendix A: Activity report

Initial setup and early testing (Oct–Dec 2024)

Project launch: Following the October 2024 discussion paper, the project began with exploratory testing of GenAI on the family law emergency motion pathway. This pathway was chosen because affidavit narrative quality is critical in urgent family matters and because it offered structured inputs suitable for GenAI experiments.

Initial experiments: Initial tests were conducted in an OpenAI API Playground environment, using an anonymized emergency motion question-and-answer set as input. GenAI was first asked to draft an affidavit based on the input and then to evaluate its own drafts for clarity and completeness. These early runs played with different models, model configuration settings (such as temperature and maximum output length), and variations in system instructions. This provided the team with a baseline understanding of how the GenAI behaviour shifted under different conditions.

Team expansion: Additional testers joined in December, receiving training on early testing methods.

Structured experimentation (Jan–April 2025)

Controlled testing: Structured experiments continued using the same anonymized emergency motion scenario as a stable baseline. The team adjusted one variable at a time (i.e. input format [Q&A versus benchmark affidavit], prompt design, model type, temperature, etc.) to assess their impact on the GenAI output.

Synthetic data: Work also began on assembling an initial library of anonymized emergency motion affidavits, where identifying details were removed and narrative depth sufficient for repeatable testing was retained.

Prototype integration with Langfuse: Tangowork consultants connected the Narrative Assistant prototype environment to Langfuse, an open-source

observability platform for AI systems. Langfuse automatically logs each test run, capturing the prompt, the AI's output, and the associated LLM evaluation. This integration was significant because it created transparency, enabled systematic evaluation, and supported scalability.

Benchmarking process: A formal benchmarking framework was introduced. It included a line-by-line comparison of GenAI narratives against benchmark (baseline) affidavits, followed by scoring across five categories: structure, accuracy, argumentative strength, clarity, and omission of relevant details.

Early expansion: Once stable conditions for testing had been established, the team expanded benchmarking and evaluation beyond the original scenario, using a small number of additional anonymized affidavits.

Suggestions: The team also began exploring how GenAI could provide suggestions in the form of follow-up questions. After an initial draft was generated by GenAI, users would be prompted to add or clarify details in their affidavits. For example, a suggestion might ask a user to explain when a key event occurred or to expand on how a particular outcome affected them. The goal was not to change the draft itself, but to provide users with a structured way to edit and improve the GenAI-generated affidavit by adding more comprehensive information. Towards the very end of the period, the team began to introduce legal reasoning into the suggestion prompting.

Expansion and iteration (May–June 2025)

Team growth and calibration: In May, six new testers joined the project, all of whom were law students. After being trained in the benchmarking and evaluation framework, weekly calibration sessions quickly became a central practice. These sessions ensured the reliability of human scoring with reduced subjectivity and also provided a forum for the team to refine and strengthen the evaluation rubric itself.

Broader testing scope: With the expanded team, testing extended from the emergency motion to the Small Claims Court pathway, beginning with claims for unpaid services. To prepare for this, members of the project team created anonymized Small Claims Court benchmark pleadings that were structured to provide realistic test inputs without exposing private details. Pre-pathway questionnaires were also developed to capture additional context not collected by the generic CLEO Guided Pathway interview questions.

Guided pathways input testing: In addition to the experiments with anonymized benchmark affidavits and pleadings, testers also entered hypothetical scenarios directly into the relevant Guided Pathways to generate an emergency motion

affidavit or a Small Claims Court pleading. The outputs were then run through the Narrative Assistant and evaluated using the established framework. Some testers intentionally replicated conditions that real users might experience. For example, they added spelling or grammatical errors to simulate users whose first language is not English. Sometimes testers left questions unanswered or marked “I don’t know” to mirror incomplete or uncertain responses. This approach ensured that the testing process captured not only controlled test conditions but also the variability and challenges of genuine user input.

Prompt refinement: Building on family law emergency motion experiments, the team introduced a new prompt for affidavit drafting that embedded legal reasoning.

Suggestions: Expanding on the work from the initial suggestion phase, the follow-up suggestion system was formalized into a three-part Suggestion Framework (Claim Checker, Story Checker, and Evidence Checker) to generate safe, targeted prompts encouraging users to strengthen their narratives.

Evaluation and transparency enhancements (July–Aug 2025)

LLM evaluation priority: Strengthening LLM evaluation became a central focus during this period. The project team enhanced LLM evaluation with revised prompting for each of the five rubric criteria, aiming to narrow the gap between AI and human scoring. The team prioritized this work to ensure that LLM evaluation could become more consistent, systematic, and scalable, such that it could be relied upon for future testing as well as implementation of the Narrative Assistant. Despite the great strides made on this front, improving LLM evaluation continues to be ongoing and evolving work.

Experimentation expansion: The team expanded experimentation to include personal injury claims in Small Claims Court, alongside continued work on existing case types.

Highlight Changes feature: With support from Tangowork, the team launched a new feature in the prototype that automatically identifies relevant GenAI edits, additions, or omissions in a draft, and provides notes that explain each change. The feature shows users exactly how their text was modified and reduces the risk of discrepancies between the user’s intended meaning and the edited version of the affidavit. It provides greater transparency and avoids undermining the user’s credibility.

Suggestions review: The Claim Checker and Story Checker outputs were compiled and reviewed to confirm their safety and relevance across multiple test runs. The Evidence Checker suggestions feature will be developed at a later time.

Expanding pathway coverage: The project team developed preliminary narrative prompts, follow-up suggestions, and pre-pathway questionnaires for additional Small Claims Court types and for a new peace bond pathway. Each prompt was infused with legal reasoning.

Chatbot development: Work began on a prototype chatbot to provide real-time technical support within the Guided Pathways. The team created a content management system, curated material from CLEO's public legal information resources, and launched an internal preview for testing common user queries.

Appendix B: Scoring for benchmarking framework

	0 Unusable	0.25 Needs Improvement	0.5 Acceptable	0.75 Accurate	1 High Quality
Structure	Missing sections, no logical structure	Sections present but out of order, or the flow of information is otherwise obstructed	All sections included and ordered in alignment with the benchmark	Well-structured affidavit with improved flow of information	Fully developed affidavit with no missing elements, follows ideal logic
Accuracy	3+ major inconsistencies (contradictions, false statements)	1-2 major inconsistencies (incorrect names or fact patterns)	Minor inconsistencies that do not alter the case facts	All claims substantiated and aligned with benchmark	Perfect accuracy, no inconsistencies
Argumentative Strength	Lacks relevant detail, weak reasoning	Some incidents mentioned but underdeveloped	Mostly well-detailed but missing minor context	Well-written with strong supporting details	Exceptionally detailed and persuasive
Clarity	Contains incoherent or incomplete sentences, much less clear than the benchmark	Disjointed, overly complex wording reduces readability	Generally clear but has 1-2 complex or awkward sections	Clear, concise language with no unnecessary complexity	Exceptionally well-written, fully readable
Omissions of Relevant Details	Multiple missing key facts (name, date, location) or major contextual gaps	1 key fact missing, unclear, or a major contextual gap	All key facts present but minor contextual gaps exist	All relevant details included with proper explanation	Perfect alignment with the benchmark