Negotiations MCP

Making climate negotiation evidence more accessible

GYC AI Labs | White paper | Version 0.6 | 2 November 2026

Language revision of the 20 October 2026 text. The service figures in section 4 are the status read from that date.

Executive overview

GYC AI Labs is building the Negotiations MCP so youth delegates and civil society can connect a priority to the public record of climate negotiations. A person can ask a compatible AI assistant about an issue or a position, look through the related discussions and documents, and follow the evidence toward questions, proposals or next steps.

The difficult work sits between finding a document and seeing how it matters. A participant needs to establish who said what, whether a text is adopted or proposed, which body is responsible, and where a priority might fit. The project treats that work as a design problem: make the evidence easier to navigate, and leave interpretation, representation and political decisions with people.

The service prepares public documents for retrieval and exposes them through the Model Context Protocol (MCP). A connected assistant can search the collection and read preserved source versions with page and passage references. The assistant writes the explanation in the user's own application. The evidence service supplies the source material and its context. MCP is the standard that connects them. [1, 4, 5]

Preparing passages once lets a later question retrieve a relevant extract from the index. That can reduce input tokens. The size of any saving depends on retrieval quality, follow-up calls and how much context the question needs. This paper reports no measured token-saving rate and no participant outcome. [5, 7]

The contribution we want to make

The service we are building combines usable access, traceable sources and negotiation-specific context. It should take someone from a concern, such as meaningful participation, to the related issues, institutions, documented positions and possible routes through the process.

This paper describes the use case, the technical approach, the operating boundaries and the proposed evaluation. It uses an authenticated service-status read from 20 October 2026, project design and release records, and the primary documentation on MCP and retrieval-augmented generation. Design intentions and proposed studies are labeled as such. Observations describe the service as it stood on that date. [1-8]

Central research question: Can this approach reduce the effort and model context needed to prepare a useful, source-supported negotiation brief while preserving the user's political intent and the ability to verify claims?

Navigate: Use case | Architecture | Token efficiency | Evaluation | Innovations | References | Project evidence

1 The participation gap and practical use case

Access requires more than document availability

The UNFCCC document portal already provides access to documents, decisions and topic filters. The gap addressed here is the additional work needed to connect those records to a participant's question. A relevant passage may sit inside a report, refer to an earlier decision or concern a body whose role is unclear to the reader. [3]

Our working premise is that this burden matters for participants with limited research support, including youth delegates early in their negotiating experience. The initial research plan prioritizes topic-informed delegates and separates subject knowledge from experience navigating the process. [6]

A question-led interaction

Consider a delegate working on meaningful youth participation. They could ask their connected assistant:

> I am working on meaningful youth participation. Using public negotiation records, identify related discussions and proposals. Show who is involved, the status and date of each source, the relevant passages, and what remains uncertain. Explain possible connections to my priority without assuming that anyone has endorsed my position.

In this illustration, the assistant would first establish the available evidence and its limits, search for relevant material and read enough surrounding context to explain the result. It could then help the user refine the question, compare documented approaches or investigate the mandate of a relevant body. [5, 6]

What a useful response should provide

A useful response should explain the issue in accessible language, attribute positions to the actors who expressed them, and identify the type, date and status of each source. It should connect factual claims to inspectable passages and distinguish those claims from the assistant's interpretation or proposed wording.

The user should be able to ask follow-up questions: Which text supports this? Is it a decision or a submission? What earlier instrument does it reference? What evidence is missing? A possible procedural route should be presented with the evidence for its mandate and with any unresolved timing or eligibility questions. [5, 6]

What the negotiation record can show

Public documents can show documented issues, proposals, mandates and outcomes. They are a partial view of what is discussed in negotiation rooms. The service covers the public collection it holds. Live rooms, informal exchanges and unpublished positions sit outside it. A missing search result leaves open the possibility that the issue was discussed somewhere else. Answers need an explicit source window, and they should point the user to developments beyond that window. [4]

2 How the evidence service works

A shared connection to prepared evidence

MCP is an open standard for connecting AI applications to external information and tools. The Negotiations MCP uses this interface to make a curated evidence collection available to compatible assistants. The protocol supplies a common connection; answer quality also depends on the client, model, prompts, retrieval and source coverage. Compatibility therefore needs to be verified in each intended client. [1, 6]

Retrieval-augmented generation is the technical precedent for selecting external passages before a model writes an answer. The original RAG research combines retrieval with generation for knowledge-intensive tasks. The Negotiations MCP uses a related retrieval principle through a tool interface. The training method and the benchmark results stay with that paper. [2]

From source document to cited passage

1. Admit and identify a source. Record its origin, title, status and available dates. Admission is a product decision about whether to keep the source. Authority and interpretation stay with the document and the reader.

2. Preserve the version. Retain original source bytes and identify the extracted version so later citations can refer to the material actually used. A later document should not silently change the meaning of an earlier citation.

3. Extract and organize the text. Divide extractor output into locatable blocks with page references and document metadata. Blocks are navigation aids. They can differ from the officially numbered paragraphs of a negotiating text. Tables, footnotes, scanned pages and reading order can require a look at the original PDF.

4. Retrieve a bounded evidence set. Search returns candidate passages. Lexical matching can surface specific wording. Semantic retrieval can help with related language where that index is available. A high search rank marks relevance. Claim support and current negotiating status come from the source text.

5. Read context and explain. The assistant reads the cited version and additional context where needed, then constructs an answer with sources, interpretation and gaps visible. The user can inspect the evidence and challenge the explanation. [4, 5]

The diagrams below separate three things: the released service, a reference client interaction, and the planned intake loop. [4, 5, 8, 9]

2A Service infrastructure and access

The documented deployment separates the HTTP evidence service from protected contribution and operational state. Ordinary readers receive public evidence tools; additional scopes expose contribution, review and bounded administration functions. A separate worker handles explicitly requested, bounded jobs. [8, 9]

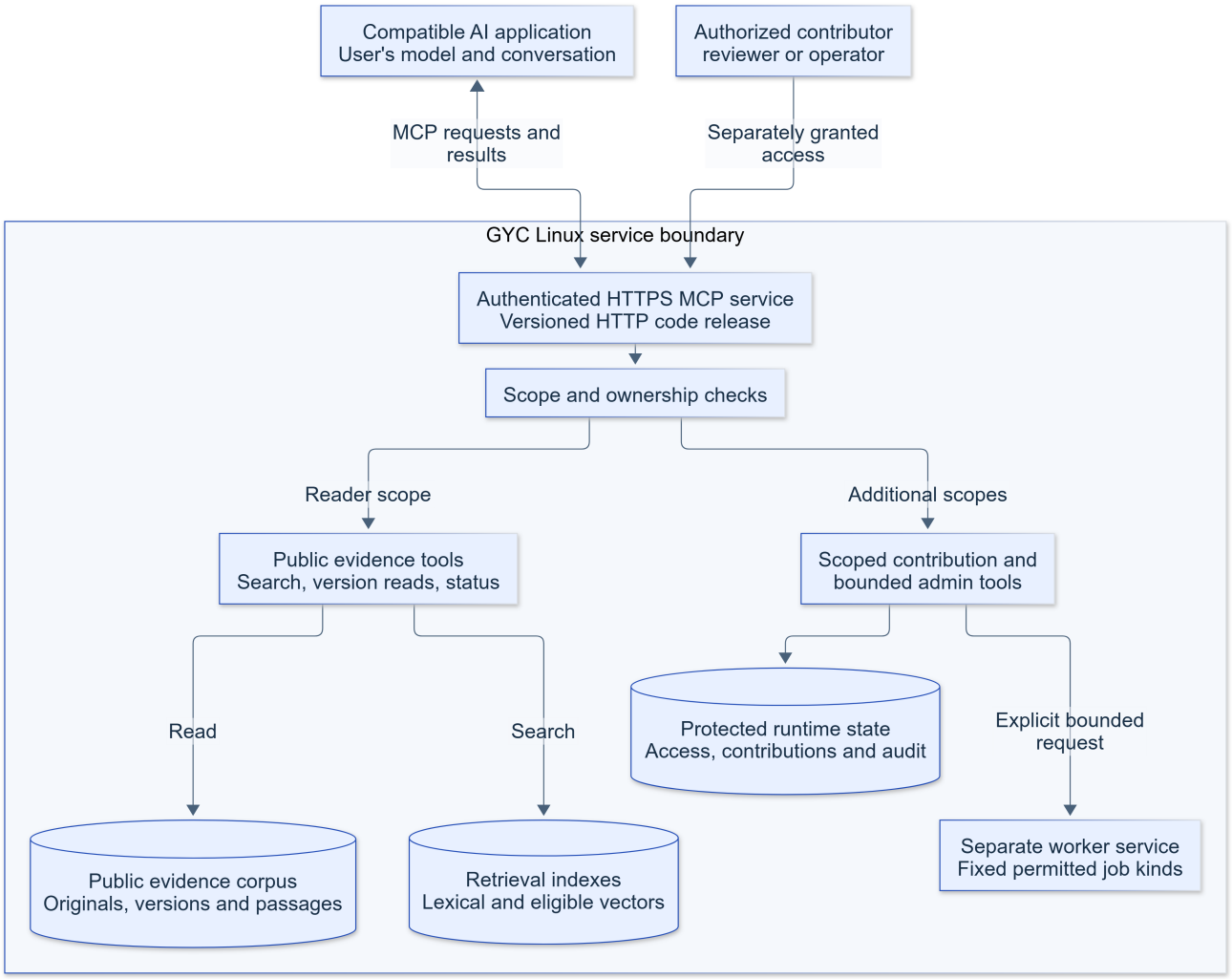


Figure 1. Documented service architecture. The corpus and indexes support retrieval. Protected runtime state has its own role and access policy. The user's model runs in the connected AI application. The HTTP service and the worker are supervised separately. [8, 9]

The evidence service checks the caller's scope and the eligibility of returned material. Reader access covers public evidence. Contribution, publication and job execution use separate grants. The worker runs explicitly requested jobs of fixed kinds. It does not watch the whole corpus continuously. [8]

2B From prompt to a cited answer

A user's agent chooses how to call the evidence tools. The sequence below is a reference workflow for a compatible client. The tools return evidence and metadata. The assistant writes the explanation. [1, 4, 6]

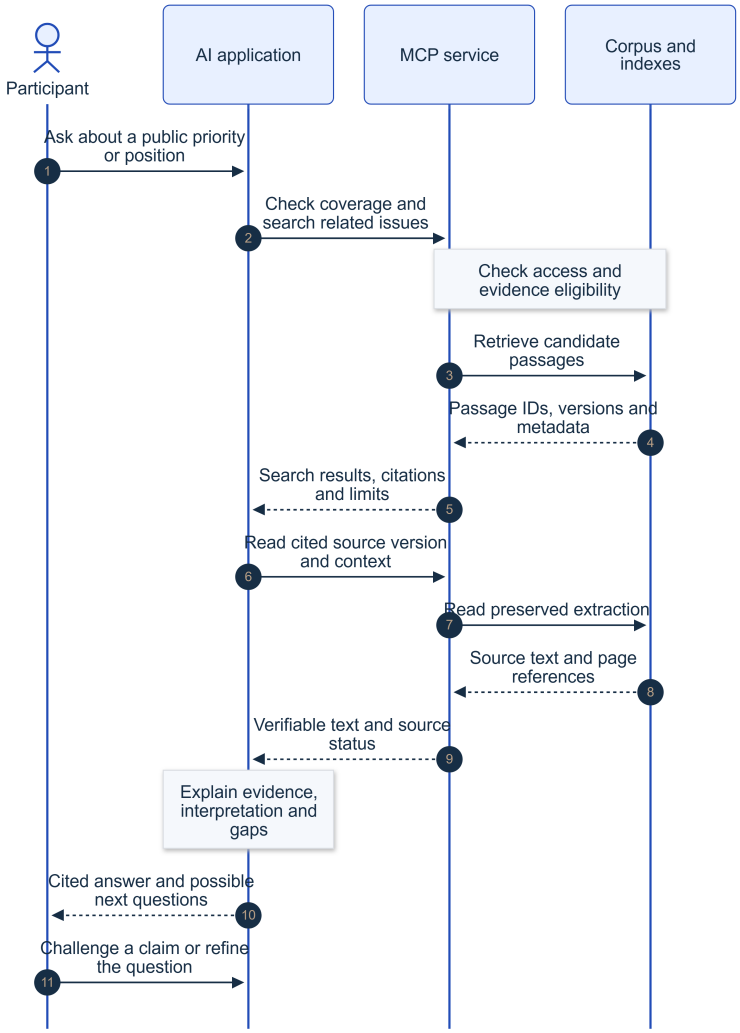


Figure 2. Reference query sequence. Search returns candidate passages. The following read supplies the surrounding text. Further searches and reads belong in the task's token and latency account. [5, 7]

The user can inspect the references and decide whether the explanation reflects their priorities. The public record can show a procedural connection. Current room dynamics, a mandate to represent a constituency, and political endorsement require other evidence. Confidential strategy stays out of this public-evidence workflow. [4, 6]

2C Planned evidence maintenance loop

The development design connects discovery, source review, extraction, retrieval and corrections. Figure 3 shows that intended loop. [5, 9]

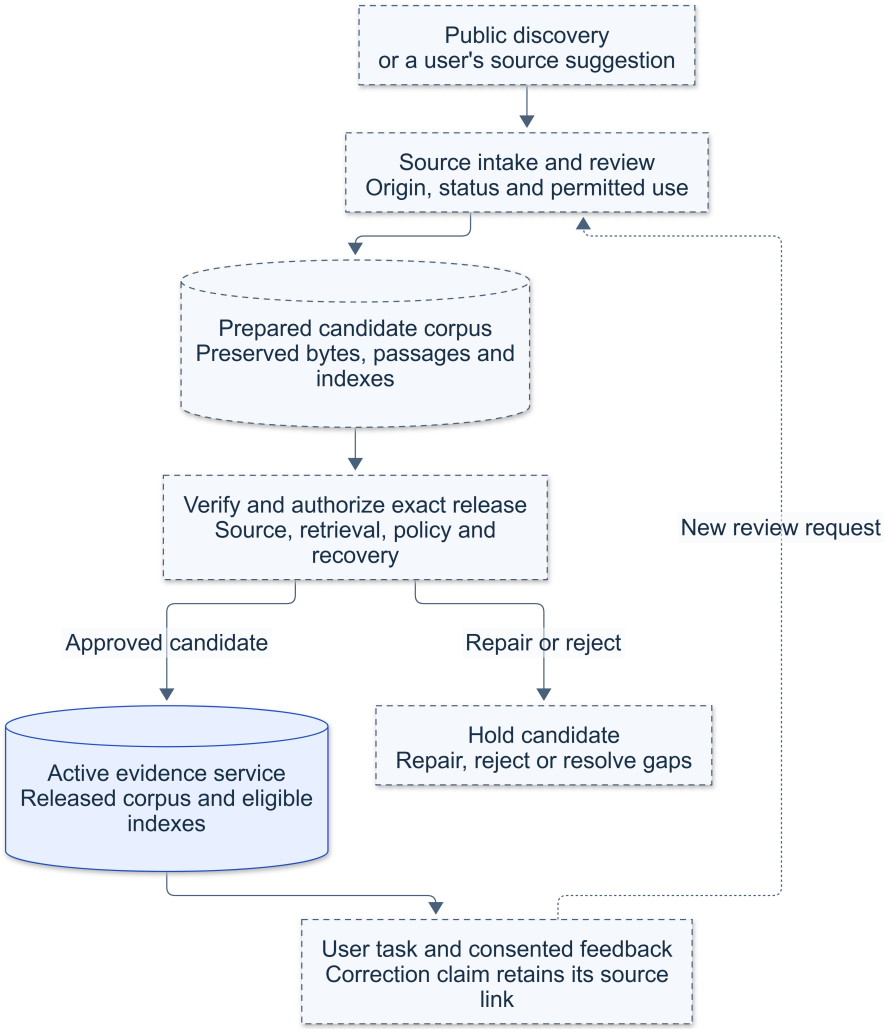


Figure 3. Intended integrated loop. Dashed borders mark the planned path. Release has its own authorization. A correction or a source suggestion stays out of the active collection until that review, and it leaves the official document unchanged. [5]

Publication and learning are separate decisions. A reviewed contribution can move through internal application and publication staging and still remain off the public service. Confidential federation and collective training sit outside this diagram and outside the current public-evidence service. [4, 8]

3 Preprocessing and token efficiency

Why preparation can reduce repeated context

Without a retrieval service, a user or agent may attach full PDFs or paste long extracts into a conversation again and again. Preparing documents once into indexed passages lets a later question request a smaller evidence set, with the metadata needed to interpret it. Less irrelevant material enters the model's prompt. [5]

The saving comes from choosing context, reusing prepared material and skipping repeated extraction. MCP itself only carries the request and the result. A token is a unit of text a language model processes. A smaller input can reduce the work of one request, depending on the model and the service. Whether the whole task becomes cheaper or faster is a measurement.

A useful comparison

For a matched task, compare the total input tokens sent across all model calls in two workflows: a baseline using complete relevant documents and a workflow using selected passages. Count tool instructions, tool results, follow-up reads, retries and repeated context under the same accounting method. Report output tokens separately, since answer length can otherwise obscure the effect of retrieval.

The token reduction would be calculated as **1 - (retrieval-workflow input tokens / baseline input tokens)**. A negative result means the retrieval workflow used more tokens. Use the same source set, model, task and answer-quality requirements. The formula is the measurement to run. This paper does not report a result for it. [7]

Efficiency depends on answer quality

Short context can omit a qualification, a neighboring paragraph or a contrasting source. When the question needs it, the workflow should retrieve more material, especially for institutional mandates, quantities, attributed positions and the status of a text. A smaller token count is useful when the answer still carries those qualifications.

Repeated searches, weak passage selection and extensive expert repair can use up the saving. Evaluation should report token use together with source coverage, unsupported claims, completion time and reviewer effort. Where quality differs, report the difference. [6, 7]

Costs beyond the prompt

Preparation has costs of its own: acquisition, extraction, review, indexing, storage and index maintenance. Semantic retrieval can add embedding costs. These should be separated from the marginal cost of a user task. Cache reuse and repeated use of the same source may improve the balance, but no break-even point or environmental benefit is established here. The finance evaluation plan already calls for token, latency, cache, retry and resource accounting; the white paper proposes extending that accounting to this comparison. [7]

4 Source integrity and human responsibility

Keep the status of evidence visible

An adopted decision, a draft, a report and a public submission make different kinds of claims. The service keeps document status and source identity so an assistant can explain the difference. A submission remains its author's position after it enters the collection. A citation shows where the text was found. Whether the assistant's conclusion follows from that text is a separate check. [4, 5]

Dates answer different questions: when a text was published, when the service collected it, and which session it belongs to. The latest fetched version has a date. Later developments need their own check. The service describes its coverage as bounded and incomplete. [4]

What human review means in practice

For a proposed briefing or recommendation, the person bringing the priority checks that it still expresses what they mean. A reviewer with relevant knowledge checks that the evidence supports its claims and that the proposed route fits the institution and process. Questions about interpretation, disagreement or missing material remain visible.

This adds time and requires scarce expertise. It can also catch errors that a fluent answer makes difficult to notice. Evaluation should count this effort, including cases where work moves from the participant to a reviewer. The score covers evidence and procedural fit. Agreement with the assistant's politics stays outside it. [6, 7]

Permissions and improvement

Ordinary reader access is separate from contribution and administration. The deployed contribution path has distinct review, application and publication-staging steps. Staging keeps an artifact internal until a later decision serves it. Changes to evidence and methods should keep source ownership and an auditable review record. [8]

The 20 October status read describes a public-sources-only service. Confidential delegation intake and collective training were off. A person's prompt and strategy stay off that public collection. This workflow is for public questions. Confidential processing would need its own authorized design for access, providers, retention and derived outputs. [4, 5]

What is established at this stage

The authenticated status read reports a public-evidence pilot with 96 document entries and 1,064 indexed vector blocks, and it reports vector readiness. Entries can be different views of one source, so the counts are an inventory of entries and blocks. They are not a count of unique PDFs, a completeness score, or a measure of search quality. The last new-version fetch recorded there is 18 October 2026. Later corpus growth is outside this paper. [4]

The path from a user's priority to an independently reviewed recommendation is still a development and validation objective. Delegate effectiveness, measured token savings and continuous coverage of the negotiations remain to be shown. [5-7]

5 Proposed research and evaluation

Separate the questions being tested

The evaluation should separate three questions: whether the evidence is sufficient, whether the explanation is accurate, and whether a person can finish the task. ALCE scores answer correctness and citation quality. RAGAs separates retrieval from generation. Both are methods to borrow. Performance figures for this project, where this paper gives any, come from the 20 October status read and the project records. Domain review also checks mandates, source status and procedural fit. Resource use is recorded beside quality. [6, 7, 12, 13]

A staged study

First, use a frozen corpus and task set. Independent domain reviewers should identify relevant passages, source status, required context and prohibited inferences before scoring the evaluated outputs. Include answerable, partially answerable, contradictory and missing-evidence questions. Keep development cases separate from held-out cases, and record client, model, prompt, corpus and retrieval versions. [7]

Second, compare workflows on matched questions. A controlled technical comparison can use complete-document context versus passage retrieval on the same source set. A separate practical comparison can examine participants' usual research method versus MCP-assisted research. These answer different questions: controlled token efficiency and usefulness in realistic preparation. Counterbalance task order and use equivalent tasks to limit learning effects.

Third, conduct a small formative study with topic-informed delegates. The existing research plan proposes initial sessions with three to five people. That size can show connection problems, confusing answers and verification barriers. A population claim needs a larger study. Record facilitator help and participant experience, and use the findings to design that study. [6, 7]

Question	Measures to record
Is the evidence sufficient?	Known-reference passage retrieval, required source pairs, coverage gaps and wrong-version results.
Are the claims supported?	Citation validity, support for substantive claims, attribution and status errors, critical misconceptions.
Does it help the participant?	Task completion, time to a usable brief, ability to verify a claim, clarity and edits needed.
What does it cost?	Total input and output tokens, tool calls, retries, latency, preparation costs and reviewer minutes.

Reporting and consent

Report raw task and participant counts, per-case failures and uncertainty. A critical error in a mandate, an attributed position or a claimed agreement blocks the affected output and stays visible in the case record. Missing evidence is scored on whether the partial answer says what is missing. [7]

Participation, collection of research notes, and reuse for evaluation or training need separate, understandable choices. Use public or synthetic task descriptions and keep little personal material. Training consent is its own choice. Praise and ordinary use do not supply it. Independent review scores evidence and procedural fit, and it leaves political preference out of the score. [5-7]

5A Proposed evaluation system

The evaluation system would sit alongside the evidence service and client workflow. It would run repeatable tasks against a fixed, permitted corpus snapshot and produce an auditable report before a candidate capability is considered for release. This is a proposed architecture, extending the existing evaluation plan. [5, 7]

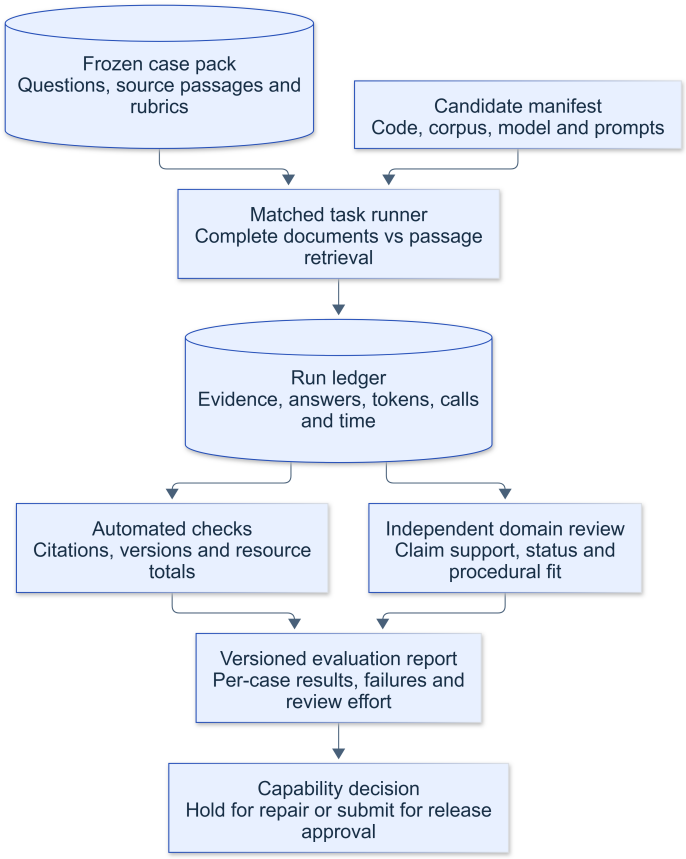


Figure 4. Proposed evaluation infrastructure. Automated checks and domain review do different work. A resolvable citation can be checked mechanically. Whether it supports a claim usually needs a person who knows the context. The report goes to a separate release decision. Publishing the report leaves the candidate where it is. [7]

Start with independently verified cases from the existing finance evaluation plan. Keep development questions separate from held-out questions. Run the same model and the same public questions with a complete-document baseline and with MCP retrieval. Where the index supports it, compare lexical and hybrid retrieval as separate conditions. Record every model call and tool result. If a token count is unavailable, record it as missing.

Store a run manifest, per-case results, a claim ledger and a resource receipt. A small report view can show retrieval relevance, unsupported claims, critical errors, total task tokens and reviewer minutes. Reviewers set acceptance criteria before scoring; critical attribution or mandate errors hold the affected capability for repair. Participant usefulness remains a separate study. [6, 7]

6 Proposed innovations

Three proposed extensions would tie the evidence service more closely to a participant's ongoing work: track how a source change affects a brief, check whether drafting preserves the user's intent, and match retrieval effort to the evidence a question needs. They build on existing project directions and published research. They are development hypotheses. [5, 10, 11]

Track how new evidence affects an existing brief

A delegate could ask, "Which parts of my brief need checking after this new document?" The proposed system would connect each factual claim and analytical dependency to its cited source version. A verified change would identify affected passages and claims, showing old and new text and the reason for review. Similar wording in a later document would not by itself establish that it replaces an earlier instrument.

This extends planned version comparison into maintenance of the user's work. It would mark an affected claim for review and leave the brief unchanged until a person accepts the update. Evaluate it on independently reviewed source pairs, including formatting-only changes. Measure missed substantive impacts, unnecessary alerts and reviewer time against a document-diff baseline. Include cases where only a claim's source status or qualification changes. [5, 7]

Check whether proposed wording preserves intent

Before drafting, a user could confirm a short statement of their public priority, intended beneficiaries and constraints. The assistant would link each proposed option to that statement and flag additions, omissions or changes in meaning. For example, it should flag a shift from a demand for decision-making participation to an option offering consultation alone, and ask the user whether that change is acceptable.

The user decides what still matches their position. An automated check offers reasons and highlights the wording for comparison. Begin with public or synthetic positions in the current workflow. Evaluate equivalent paraphrases alongside deliberately altered drafts: measure detected changes, false alarms, owner corrections and review time. A factual or procedural improvement still leaves the political intent with the user. [5, 6]

Adapt retrieval to evidence needs and a task budget

The agent could start with a compact set of exact passages and expand it when a question requires a missing mandate, qualification or contrasting source. It would preserve source status, attribution and relevant context, then answer partially, ask for clarification or identify a gap when the evidence remains insufficient. It extends task routing using adaptive retrieval and context-sufficiency research, which still needs testing on negotiation tasks. [5, 10, 11]

Compare it with fixed passage retrieval and with complete-document context on the same tasks. Record total tokens, calls, latency and reviewer effort beside claim support and omitted evidence. Include unanswerable questions and cases where a further retrieval is essential. Sufficient evidence is a recorded retrieval result. A token reduction counts when the agreed quality requirements still hold. [7]

6A Connecting the proposed extensions

These extensions would share the existing source identifiers, citation records and evaluation infrastructure. A bounded prototype can use versioned records and explicit links between claims, source versions and the user's confirmed public intent. [5]

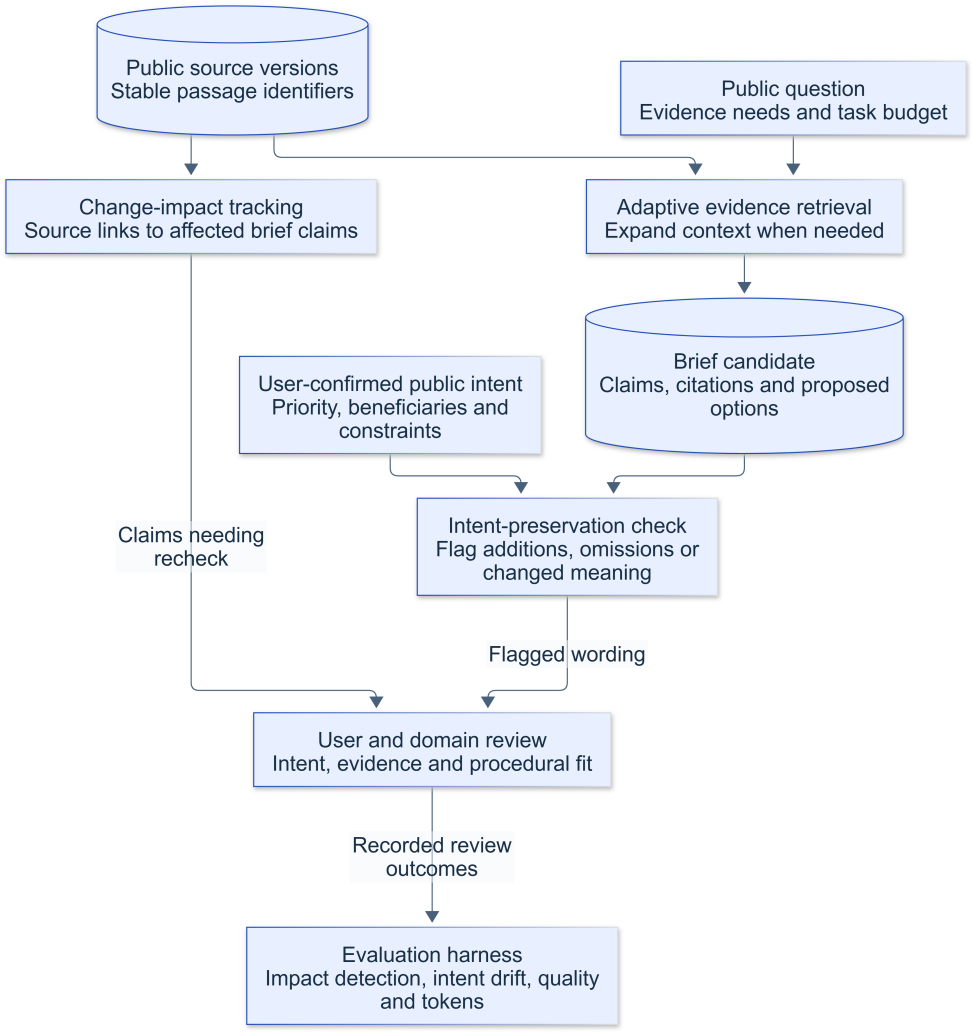


Figure 5. Proposed integration points. The same source and claim links support traceability and change-impact checks. A user-confirmed intent statement is a separate reference for drafting. The evaluation harness records retrieval and drafting runs before review, including failures a reviewer later repairs.

A first prototype should link a small public brief to its exact sources, add a reviewed source update and show affected claims. Next, test the intent check with deliberately changed wording. Add adaptive retrieval once fixed-retrieval quality and resource baselines exist. Each extension should be evaluated separately, followed by a combined task to detect interactions and duplicated effort.

7 Development and research contribution

The next evidence to establish

The immediate priority is a repeatable user task. A person connects their assistant, asks about a public priority, finds related records, verifies the relevant passages and produces a short brief with the gaps visible. Record where the user needs help and how much effort the reviewer spends correcting the result. That record is the progress measure. Document counts and tool counts are context for it. [6]

The next step is a matched efficiency comparison that holds answer quality constant and accounts for preparation costs and repeated calls. Coverage should then grow in response to important unanswered questions, with visible source dates and document status. The proposed extensions in Section 6 should retain separate evaluation cases and release decisions before they are combined into a maintained briefing workflow. [5, 7]

The aim is a shared negotiation record a person can inspect through the assistant they already use. The evaluation has to show whether preparation becomes more accessible, whether the sources stay accurate, whether political ownership stays with the user, and what the service costs to operate.

How to read the statements in this paper

Three kinds of statement appear above. Public research supplies methods. The 20 October 2026 status read and the project records in Appendix A describe the system and the design. Sections 5 and 6 are evaluation plans and development hypotheses. Their outcomes are still to be tested.

Citations use the numbers in the text. Public entries include the publisher or conference link. Internal records are unpublished. Appendix A lists each one with a file locator. The inventory in section 4 is the service's own report of what it held on 20 October 2026.

References

Public research and documentation

References retain the numbers used in the text. Internal project sources, numbered 4-9, are in Appendix A. Public metadata and links below were checked on 20 October 2026. Public studies are cited for method. Counts and dates in this paper come from the 20 October 2026 status read.

[1] Model Context Protocol. (2026). What is the Model Context Protocol? Official documentation, version 2026-07-28. Locator: opening definition and architecture overview. Cited for MCP as a connection standard.

[2] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W.-t., Rocktaschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 33. Locator: abstract. arXiv record and versions; preprint DOI. Cited for the retrieval-generation method.

[3] United Nations Framework Convention on Climate Change. (n.d.). Documents. Official document portal. Locator: document categories and topic filters. Cited for the existing document portal.

[10] Jeong, S., Baek, J., Cho, S., Hwang, S. J., and Park, J. C. (2024). Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. Proceedings of NAACL-HLT, Volume 1: Long Papers, pp. 7036-7050. Association for Computational Linguistics. DOI: 10.18653/v1/2024.naacl-long.389. Locator: abstract. Cited for adapting retrieval effort to question complexity.

[11] Joren, H., Zhang, J., Ferng, C.-S., Juan, D.-C., Taly, A., and Rashtchian, C. (2025). Sufficient Context: A New Lens on Retrieval Augmented Generation Systems. International Conference on Learning Representations. Conference paper; arXiv v3. Locator: abstract. Cited for the difference between relevant context and context sufficient to answer.

[12] Gao, T., Yen, H., Yu, J., and Chen, D. (2023). Enabling Large Language Models to Generate Text with Citations. Proceedings of EMNLP, pp. 6465-6488. Association for Computational Linguistics. DOI: 10.18653/v1/2023.emnlp-main.398. Locator: abstract and ALCE evaluation description. Cited for scoring answer correctness separately from citation quality.

[13] Es, S., James, J., Espinosa Anke, L., and Schockaert, S. (2024). RAGAs: Automated Evaluation of Retrieval Augmented Generation. Proceedings of EACL: System Demonstrations, pp. 150-158. Association for Computational Linguistics. DOI: 10.18653/v1/2024.eacl-demo.16. Locator: abstract and evaluation dimensions. Cited for retrieval and generation diagnostics.

Project information: Negotiations MCP | GYC AI Labs.

Appendix A Project evidence

Internal records below are unpublished. Each locator is a path in the GYC AI Labs project tree. The 20 October 2026 status read is the inventory used in section 4. Older runtime counts in the notes are left in those files and are not reused here.

[4] Negotiations MCP. Authenticated workspace_status response, observed 20 October 2026. Public-evidence inventory, vector readiness, freshness, confidentiality and coverage. The observation reports what the service said it held. Participant outcomes are outside it. Preserved summary: output/pdf/negotiations-mcp-white-paper-evidence.json.

[5] GYC AI Labs. Evolve Negotiations Collective Brain: design. openspec/changes/evolve-negotiations-collective-brain/design.md. Intended architecture and capability boundaries.

[6] GYC AI Labs (13 October 2026). Negotiations MCP: product gaps and youth-delegate beta plan. openspec/notes/negotiations-mcp/beta-readiness-2026-10-13.md. Research rationale and proposed sessions. The runtime inventory in that note is older than [4].

[7] GYC AI Labs (17 October 2026). Climate finance outcome evaluation. Proposed protocol, openspec/notes/negotiations-mcp/climate-finance-2026-10-17/eval-plan.md. Evaluation design. The protocol had not been run as a completed study when this paper was written.

[8] GYC AI Labs (17 October 2026). Negotiations MCP owner and admin access. openspec/notes/negotiations-mcp/owner-admin-deployment-2026-10-17.md. Scoped access and staged publication. The inventory in that note is older than [4].

[9] GYC AI Labs. Negotiations Collective Brain project map and public corpus release record. openspec/notes/negotiations-mcp/collective-brain-project-map.md and presidency-deployment-2026-10-16.md. Service separation and release boundaries. The inventories in those notes are older than [4].