

Filtrum-Safety: LLM-Agnostic RAG for Reducing Legal Hallucinations in Civil-Law Contexts

Ramirez Andrade Uriel Ismael Guadalupe | Peña Martínez Diego David | Cadena Alfaro Tadeo

With Apart Research

Research conducted at The Global South AI Safety Hackathon, June 2026

Abstract

The use of large language models in legal contexts can produce fabricated citations, jurisdictional confusion, and the transfer of Common Law categories to civil-law problems. This risk increases when students or practitioners query an LLM without an external method for source verification or when free versions limit the number of documents that can be uploaded. We developed Filtrum-Safety, an LLM-agnostic Retrieval-Augmented Generation (RAG) tool that receives a variable number of legal documents, extracts and cleans their content, splits it into chunks, generates embeddings, and retrieves the most relevant fragments using cosine similarity. The user defines a query, Top K, and a minimum similarity threshold; the system exports a Markdown file with selected fragments, sources, and similarity scores, ready to be used as context in the preferred LLM. In a preliminary test with three Mexican legal documents and 427 indexed chunks, Filtrum-Safety produced compact and traceable context for a query about an amparo claim based on omission of search duties. The system does not eliminate hallucinations, but it reduces their surface by separating retrieval from generation. Applying the five-criterion evaluation matrix to the same $n=1$ case produced a higher preliminary score for the Markdown-assisted condition (10/10) than for the direct LLM condition (5/10), suggesting that retrieved context can improve document grounding, verifiability, and jurisdictional control in legal generation tasks.

Keywords: Retrieval-Augmented Generation (RAG); document retrieval; embeddings; cosine similarity; legal hallucinations; civil law; AI safety; contextual Markdown.

Summary

The use of language models in law can produce fabricated citations, jurisdictional confusion, and the transfer of Common Law categories into civil-law legal problems. This risk increases when students or professionals consult an LLM without an external source-verification method or when free versions limit how many documents can be uploaded. We developed Filtrum-Safety, an LLM-agnostic RAG tool that receives a variable number of legal documents, extracts and cleans their content, divides it into chunks, generates embeddings, and retrieves the most relevant fragments through cosine similarity. The user defines a query, a Top K, and a minimum similarity threshold; the system exports a Markdown file with fragments, sources, and scores, ready to be used as context in the preferred LLM. In a preliminary test with three Mexican legal documents and 427 indexed chunks, Filtrum-Safety built compact and traceable context for a query about an amparo claim based on omission of search duties. The system does not eliminate hallucinations, but it reduces their surface by separating document retrieval from generation. When the five-criterion matrix was applied to the same $n=1$ case, the Markdown-assisted condition obtained a higher preliminary score (10/10) than the direct LLM response (5/10), suggesting that retrieved context can improve document grounding, verifiability, and jurisdictional control in legal generation tasks.

Keywords: Retrieval-Augmented Generation (RAG); document retrieval; embeddings; cosine similarity; legal hallucinations; civil law; AI safety; contextual Markdown.

1. Introduction

Large language models (LLMs) are increasingly used to draft, summarize, search doctrine, and prepare legal arguments. In Mexico and much of Latin America, including South America, this adoption takes place within predominantly civil-law systems: Mexico's legal system stems from the civil-law tradition, and Latin American jurisdictions share a civil-law heritage while preserving national variation. In this context, statutory authority, case law, and doctrine are not always organized or interpreted in the same way as in Common Law jurisdictions. The central risk is not only that a model may answer incorrectly, but that it may do so with the appearance of legal certainty: citing nonexistent rules, mixing jurisdictions, importing foreign categories, or presenting reasoning that cannot be verified.

This project addresses that failure mode from an applied AI safety perspective: reducing dependence on the LLM's internal knowledge and requiring generation to start from documentary evidence provided by the user. The threat model is a user who directly asks an LLM about a local legal problem and receives a plausible response that is not grounded in the relevant sources of their jurisdiction.

Filtrum-Safety proposes a prior document-retrieval layer. Instead of sending all files to the LLM, the system processes “ n ” documents, identifies the fragments closest to the user query, and exports a compact, auditable, and reusable Markdown file. The practical contribution is twofold: it enables users to work with larger document collections than some free LLM upload limits allow, and it improves traceability by providing the model only with retrieved fragments that include source and similarity score.

Although the validation case focuses on law, the architecture has a multidomain scope. Any document-intensive area - health, education, compliance, engineering, public policy, or knowledge management - can benefit from a layer that retrieves evidence before generating a response. Our main contributions are: (i) a functional LLM-agnostic RAG prototype for Spanish-language legal documents; (ii) a Markdown export flow that separates retrieval from generation; and (iii) a preliminary qualitative evaluation centered on document grounding, jurisdictional confusion, and doctrinal transfer.

2. Related Work

RAG combines a model's parametric memory with explicit retrieval from an external index, which can improve factuality, specificity, and updatability in knowledge-intensive tasks [1]. In law, this approach is especially relevant because the quality of a response depends on its relationship to statutory texts, case law, doctrine, and the applicable jurisdiction.

Recent literature on legal AI warns that hallucinations do not disappear simply by using more advanced models; they require mechanisms for truthfulness, traceability, and human oversight [2]. In legal retrieval, Lima proposes multilayer embeddings for legislative texts, respecting the hierarchical structure of articles, paragraphs, and chapters [3]. Guerra, Monti, and Fernandez apply RAG to Spanish-language academic documents through extraction, segmentation, vectorization, and natural-language querying [4]. Work such as LegalGraphRAG and MVRAG explores structured retrieval and multiple perspectives for legal or knowledge-dense tasks [5,6].

Filtrum-Safety differs because it does not seek to replace the LLM or become a closed legal-answering platform. Its contribution is to act as an intermediate tool: it receives user files, retrieves the most relevant fragments through cosine similarity, and generates a Markdown file transferable to any LLM. This makes it useful for hackathons, classrooms, small legal practices, and users who need to control context before generating a response.

3. Methods

The system architecture follows four phases: ingestion, indexing, retrieval, and export. During ingestion, the user uploads legal documents in formats compatible with the prototype extractor. The system extracts text, normalizes line breaks, and removes basic noise. During indexing, the text is divided into chunks that seek to preserve semantic coherence and, when the document allows it, respect legal boundaries such as articles, subsections, or sections. Each chunk is transformed into a high-dimensional vector representation through embeddings.

The prototype version used the jina-embeddings-v2-base-es embedding model, selected because of its orientation toward semantic retrieval in Spanish and its usefulness for comparing queries and document chunks within the same vector space. This model can be replaced in future versions: the architecture allows the embedding model to change without modifying the central principle of retrieval and context export.

During retrieval, the user writes a query or context. Filtrum-Safety generates the embedding for that query and computes cosine similarity against the vectors of the indexed chunks. Cosine similarity compares the orientation of two vectors and makes it possible to estimate semantic closeness between the user's need and each document chunk. The system then filters results using two configurable parameters: Top K, which defines how many matches are exported, and a minimum similarity threshold, which prevents low-affinity chunks from being included.

Finally, the system builds an output Markdown file. The Markdown is not intended to answer by itself, but to function as input context for an external LLM. The recommended structure includes: user query, search parameters, list of retrieved chunks, source file name, similarity score, and a document-grounding instruction so that the LLM answers only based on the provided context, declares the jurisdiction, and states when the evidence is insufficient.

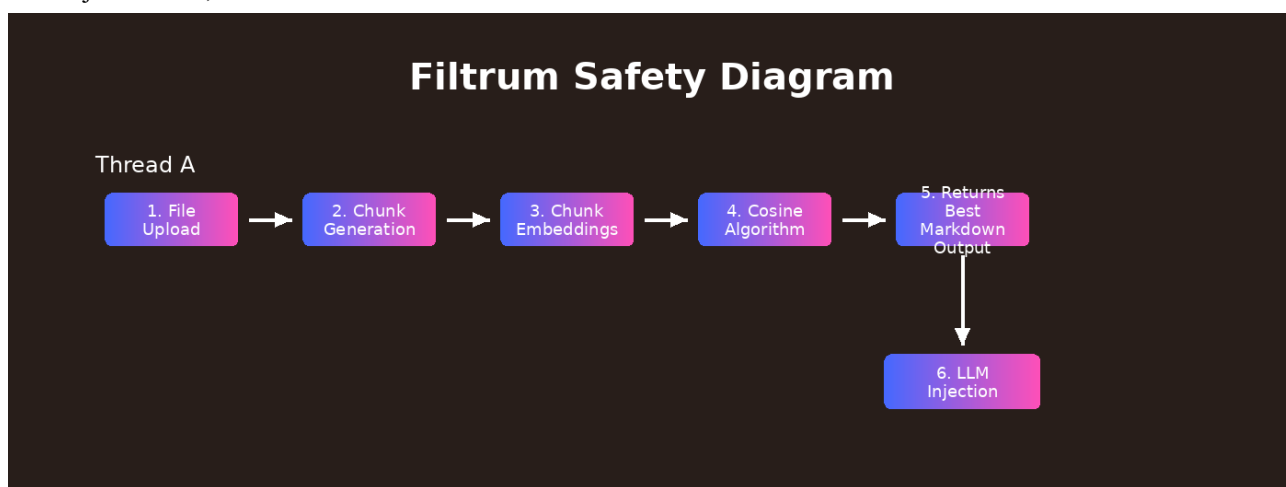


Figure 1. General architecture of Filtrum-Safety: the system separates document retrieval from generation to provide compact, traceable, and configurable Markdown context.

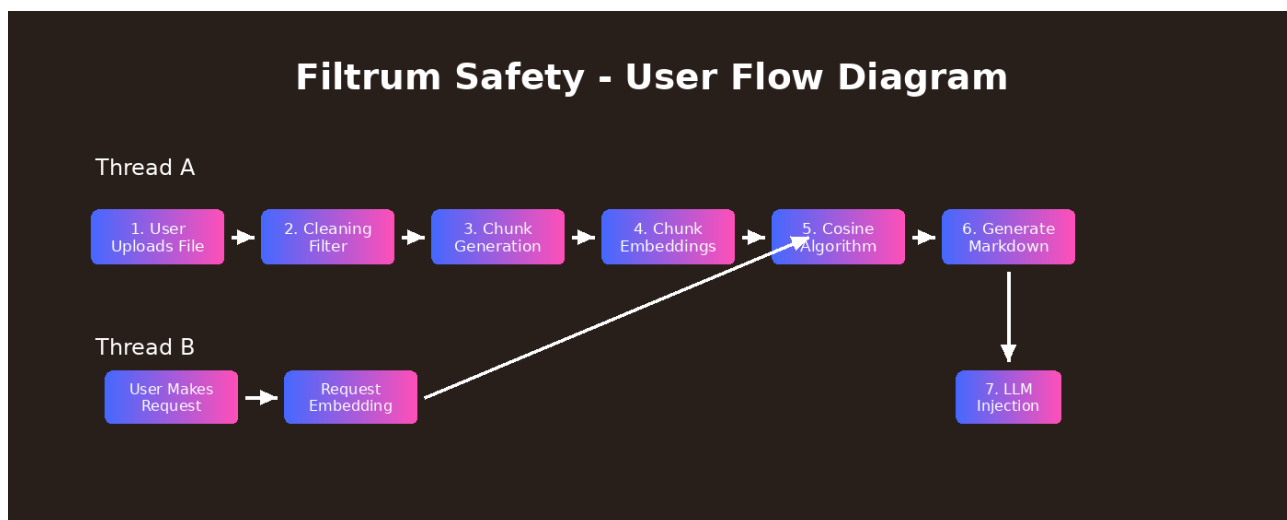


Figure 2. User flow and document retrieval: the document base is indexed into chunks and embeddings; it is compared through cosine similarity and exported as contextual Markdown for any LLM.

Preliminary prototype configuration.

Element	Design decision
Input	Variable number of documents uploaded by the user.
Segmentation	Semantic or structural chunks to preserve legal coherence.
Embedding model	jina-embeddings-v2-base-es, used to represent queries and chunks in a shared vector space.
Retrieval	Cosine similarity with configurable Top K and minimum similarity threshold.
Output	Markdown with chunks, sources, scores, and document-grounding instructions.
LLM use	The user's preferred LLM receives the Markdown as context, not the original corpus.

4. Results

During the hackathon test, we loaded a Mexican-jurisdiction legal corpus composed of three documents: (1) Ley General en Materia de Desaparición Forzada de Personas, Desaparición cometida por Particulares y del Sistema Nacional de Búsqueda de Personas, a Mexican federal law on enforced disappearance and the national search system; (2) a local Jalisco law uploaded as 03-05-21-ley general jalisco.pdf; and (3) LGP.pdf, a complementary document included in the team's test corpus. Altogether, the system generated 427 indexed chunks.

The evaluated query was: “What interpretation criteria does Article 3 of the Ley General en Materia de Desaparición Forzada de Personas require in order to support an amparo claim based on omission of search duties in Mexico?” This example was selected because it requires jurisdictional awareness: it is not enough to answer about search duties in general; the response must rely on Mexican legal sources and distinguish between the federal framework and the local Jalisco reference.

For retrieval, we used Top K = 5 and a minimum similarity threshold of 0.65. The threshold was defined as an exploratory value during the manual prototype test: it allowed legally related fragments to be retained without saturating the Markdown with weakly connected text. It is not presented as a universal value; it should be recalibrated according to corpus size, domain, and embedding model.

The evaluation was qualitative and compared two conditions for the same n=1 case: a direct LLM response without retrieved context and a response assisted by the Markdown generated by Filtrum-Safety. We do not report statistical significance because the hackathon scope allowed a preliminary validation, not a large-scale study. To turn the comparison into observable evidence, the matrix in Appendix B was applied to both outputs. The complete responses are documented in Appendix C through screenshots; the comparative score is summarized in Table 2.

Table 1. Preliminary qualitative comparison.

Criterion	Direct LLM	LLM with Filtrum-Safety Markdown
Alignment with uploaded corpus	Low or not verifiable: the model answers from its internal memory.	Higher: the response starts from retrieved chunks.
Traceability	Limited: sources and citations may not be auditable.	Improved: each chunk can preserve file, score, and context.
Hallucination risk	Higher, especially in citations or unverified institutions.	Reduced, though not eliminated; it depends on corpus quality and instruction.
Jurisdictional confusion	May mix external or general criteria.	Lower if the corpus and Markdown declare jurisdiction and source.
Cost and upload limits	Depends on the limits of the LLM used.	The corpus is processed outside the LLM and only selected context is sent.
Human oversight	Necessary, but with less visible local evidence.	Necessary and easier, because retrieved chunks are reviewable.

Table 2. Application of the qualitative matrix to the n=1 case of amparo based on omission of search duties.

Criterion	Direct LLM (0-2)	LLM with Filtrum-Safety Markdown (0-2)
Document grounding	0 - It does not rely on retrieved chunks from the uploaded corpus.	2 - The response starts from chunks retrieved and exported in the Markdown.
Verifiability	1 - It mentions legal concepts, but does not clearly locate the exact source within the corpus.	2 - It preserves file, chunk, and similarity score, facilitating human review.
Jurisdiction	1 - It locates the problem in Mexico in general terms, but does not clearly distinguish the federal/local corpus used.	2 - The query, corpus, and Markdown declare Mexican jurisdiction and make it possible to separate federal and local reference.
Legal hallucination	1 - Possible risk: the output depends on internal memory and may introduce authorities or criteria that cannot be verified.	2 - No rules external to the retrieved context were observed in the reviewed output.
Doctrinal transfer	2 - No explicit importation of Common Law categories was observed in this case.	2 - No doctrinal transfer was observed; generation remained restricted to the provided corpus.
Preliminary total	5/10	10/10

Note: these scores are preliminary and correspond to a single test case (n=1). Their purpose is to demonstrate the evaluation method; the complete outputs that justify each score are preserved in Appendix C.

Supplemental cross-model check.

To test whether the observed behavior depended on a single model, we ran a supplemental qualitative check with Claude, Gemini, and DeepSeek. The same legal query was submitted in two conditions: direct prompting without retrieved context and prompting with the Markdown context generated by Filtrum-Safety. This additional check is not treated as a statistical benchmark and does not replace the n=1 matrix in Table 2. Its purpose is to show whether the retrieval layer changes model behavior across different LLM interfaces.

Across the three models, direct prompting tended to produce broader legal reasoning based on internal or web-assisted knowledge. Markdown-assisted prompting made the provenance of the answer more explicit: the models either grounded their response in the retrieved Article 3/Jalisco context or, in the case of DeepSeek, stated that the retrieved context was insufficient to fully support an amparo-specific answer. This suggests that Filtrum-Safety can improve audibility across models even when the final generation is delegated to different LLMs.

Table 3. Supplemental cross-model qualitative check.

Model	Direct prompting without Markdown	Markdown-assisted prompting with Filtrum-Safety
Claude	Disambiguated the phrase “Article 3” using external or internal legal knowledge and introduced alternative legal sources, including the Ley de Amparo and broader disappearance-law context.	Anchored the answer to the retrieved context, especially Article 3 from the Jalisco disappearance-law corpus, making the source of the response more explicit.
Gemini	Produced broad constitutional and jurisprudential criteria, including pro persona, SCJN criteria, due diligence, and international instruments, but without a retrieved-document anchor.	Narrowed the answer to the Markdown-provided context and connected the response to the document attached as contextual evidence.
DeepSeek	Inferred general amparo requirements and possible Article 3 meanings from web/internal	Explicitly noted that the retrieved context did not fully contain amparo-specific criteria and

Model	Direct prompting without Markdown	Markdown-assisted prompting with Filtrum-Safety
	knowledge, creating a risk of answering beyond the provided corpus.	identified missing elements, which reduced overconfident extrapolation.

5. Discussion and Limitations

The preliminary results and the application of the qualitative matrix show that the retrieval layer reduces a frequent failure mode in legal queries: asking the LLM to “know” the answer without requiring it to be grounded in concrete sources. By building a Markdown file with retrieved chunks, Filtrum-Safety changes the interaction: the user not only asks a question, but also controls the document collection from which the response will be generated. This decision matters for AI safety because it promotes verifiability, reduces opacity, and facilitates human review before using a response in academic or professional contexts. The supplemental cross-model check reinforces this interpretation by showing that the benefit is not tied to a single LLM: the retrieved Markdown made the response path more auditable across Claude, Gemini, and DeepSeek.

The tool should not be interpreted as automated legal advice or as a guarantee of truth. Its main limitations are: (i) if the uploaded corpus is incomplete, outdated, or biased, the Markdown will inherit that problem; (ii) cosine similarity may retrieve semantically close but legally insufficient chunks; (iii) an external LLM may still ignore instructions or overinterpret fragments; (iv) the hackathon validation was small and qualitative; and (v) the system still requires better metadata by jurisdiction, year, source type, and normative hierarchy.

Future work should focus on multicenter evaluation, document-grounding metrics, automatic extraction of legal citations, better metadata by jurisdiction, hybrid retrieval, and instruction templates that force the LLM to declare uncertainty when the context is insufficient. In addition, future scope is not limited to law: the same architecture can be adapted to medical documents, engineering manuals, institutional regulations, public policies, educational materials, and internal corporate knowledge bases.

6. Conclusion

Filtrum-Safety demonstrates that a simple, LLM-agnostic RAG tool centered on user-provided documents can improve how generative AI is queried in civil-law legal contexts. Its greatest value is not producing a final answer, but preparing compact, traceable, and configurable Markdown context that reduces dependence on the model's internal knowledge.

Within the hackathon scope, the prototype showed technical feasibility for processing multiple documents, generating embeddings with jina-embeddings-v2-base-es, retrieving matches through cosine similarity, and exporting context usable by any LLM. The proposal does not eliminate legal hallucinations, but it offers a practical mitigation barrier: moving from answers generated from opaque memory to answers supported by reviewable documentary evidence. Its growth potential lies in every domain where the response must preserve a clear relationship with source documents.

Code and Data

Code repository: <https://github.com/urilupe/Filtrum-Safety>

Artifact/demo: https://drive.google.com/file/d/1sd4FJFMpjKo5yxzn_E-VsHb0eDfv0TiJ/view?usp=sharing

Data/corpus: legal documents uploaded during the hackathon test. When distribution restrictions exist, we recommend reporting metadata, synthetic examples, or anonymized screenshots.

References

- [1] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuettler, H., Lewis, M., Yih, W., Rocktaeschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS 2020. <https://arxiv.org/abs/2005.11401>
- [2] Dantart, A. (2025). Legal Artificial Intelligence and the Challenge of Truthfulness: Analysis of Hallucinations, RAG Optimization, and Principles for Responsible Integration. arXiv:2509.09467. <https://arxiv.org/abs/2509.09467>
- [3] Lima, J. A. de O. (2024). Unlocking Legal Knowledge with Multi-Layered Embedding-Based Retrieval. arXiv:2411.07739. <https://arxiv.org/abs/2411.07739>
- [4] Guerra, F., Monti, K., & Fernandez, J. M. (2025). Retrieval-Augmented Generation (RAG) System for Access to Academic Information. XXXI Argentine Congress of Computer Science. <https://sedici.unlp.edu.ar/handle/10915/191535>
- [5] Chen, Z., Zhang, Q., Xiang, Z., Wei, Z., Gao, L., Huang, X., Zhang, Z., & Su, J. (2026). LegalGraphRAG: Multi-Agent Graph Retrieval-Augmented Generation for Reliable Legal Reasoning. arXiv:2605.28120. <https://arxiv.org/abs/2605.28120>
- [6] Chen, G., Yu, W., & Sha, L. (2024). Unlocking Multi-View Insights in Knowledge-Dense Retrieval-Augmented Generation. arXiv:2404.12879. <https://arxiv.org/abs/2404.12879>
- [7] Avalos, F. A. (2013). Mexican Legal System. Daniel F. Cracchiolo Law Library, University of Arizona. <https://lawlibrary.arizona.edu/mexican-legal-system>
- [8] Grant, J. E. (2022). Researching the Law of Latin America. GlobaLex, New York University School of Law. https://www.nyulawglobal.org/globalex/latin_america_law_research1.html

Appendix

A. Recommended instruction for using the Markdown

Use only the context retrieved in this Markdown to answer. Indicate the jurisdiction and source whenever possible. Do not invent rules, citations, or precedents. If the context does not contain sufficient information, answer that there is not enough evidence in the provided documents and suggest which source should be reviewed.

B. Qualitative evaluation matrix

The matrix was applied to the two outputs from the same evaluated case: the direct LLM response and the Markdown-assisted response. The scale maintains three levels: 0 = deficient performance or high risk; 1 = partial performance or possible risk; 2 = adequate performance or no observed risk.

Table B1. Evaluation criteria used.

Category	Guiding question	Suggested scale
Document grounding	Is the claim supported by a retrieved chunk?	0 = no; 1 = partial; 2 = yes
Verifiability	Does the response make it possible to locate the source, document, or section?	0 = no; 1 = partial; 2 = yes
Jurisdiction	Does it clarify whether the source belongs to Mexico, Jalisco, or another jurisdiction?	0 = no; 1 = partial; 2 = yes
Legal hallucination	Does it include rules, citations, or authorities not present in the context?	0 = high; 1 = possible; 2 = not observed
Doctrinal transfer	Does it import Common Law categories without justification?	0 = high; 1 = possible; 2 = not observed

Table B2. Preliminary score applied to the n=1 case.

Category	Direct LLM	LLM with Markdown
Document grounding	0	2
Verifiability	1	2
Jurisdiction	1	2
Legal hallucination	1	2
Doctrinal transfer	2	2
Total	5/10	10/10

C. Output evidence for the n=1 case

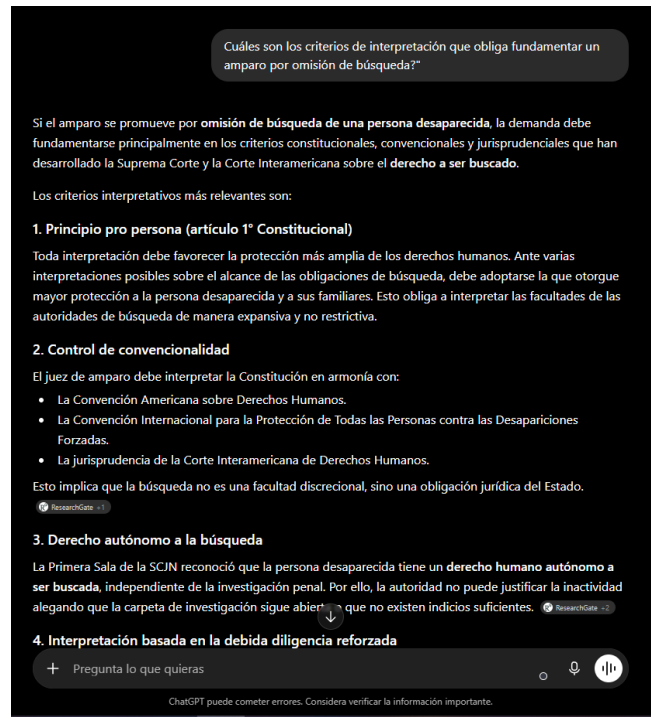


Figure 3. Direct LLM response without Markdown.

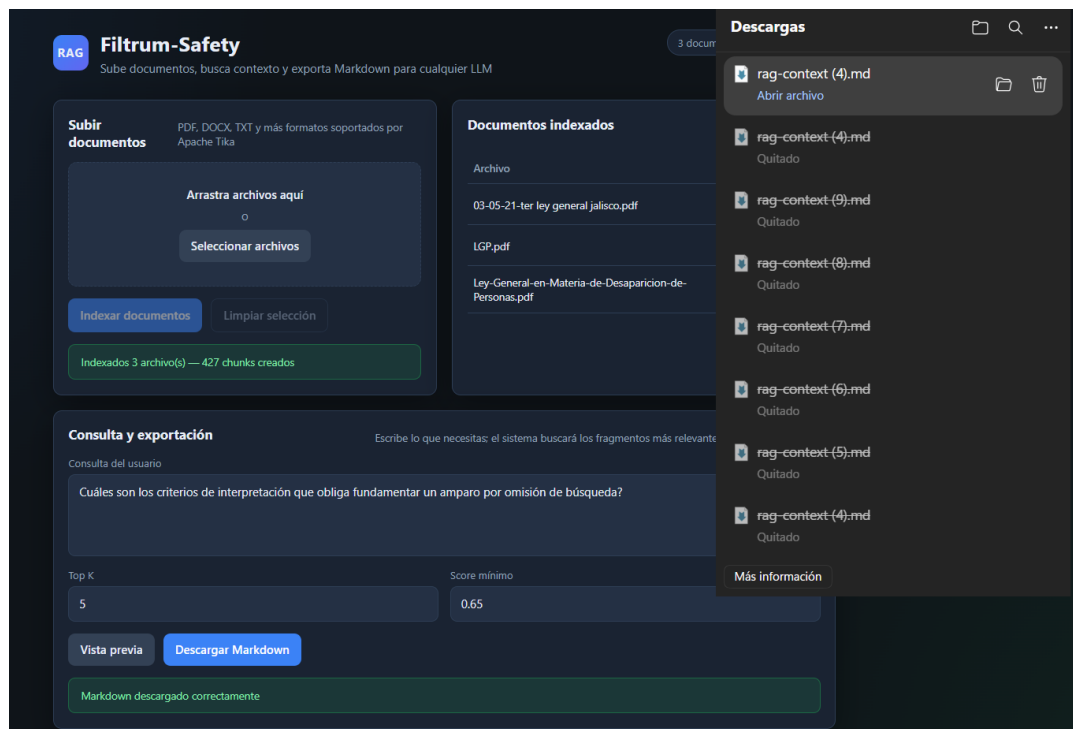


Figure 4. Filtrum-Safety interface showing Markdown generation and download with Top K = 5 and threshold = 0.65.

```
rag-context (4).md X
C:\Users\W10\Downloads> rag-context (4).md > # Contexto RAG
1 Contexto RAG
2
3 > Generado: 2026-06-21 19:57:28
4
5 ## Consulta del usuario
6
7 ¿Cuáles son los criterios de interpretación que obliga fundamentar un amparo por omisión de búsqueda?
8
9 ## Fragmentos relevantes
10
11 ### Fragmento 1 (similitud: 0.7203)
12
13 - **Archivo:** 03-05-21-ter ley general jalisco.pdf
14 - **Documento ID:** 7b4902c1-ed59-4ea6-b498-b2d04f120eac
15 - **Chunk:** 163
16
17 ```text
18 con los lineamientos correspondientes; por sí o en coordinación con
19 la Comisión de Búsqueda; XXXII. Establecer coordinación e intercambio de información constante con la Comisión de
20 Búsqueda y la Comisión Ejecutiva para la atención integral a Víctimas, a fin de evitar los
21 procesos de revictimización; XXXIII. Suministrar, sistematizar, analizar y actualizar la información de hechos y datos sobre
22 los delitos previstos en la Ley General; y XXXIV. Las demás que establezca la Ley Orgánica de la Fiscalía Estatal, el Reglamento
23 Interno de la Fiscalía Especial, la Ley General, el Protocolo Homologado de Investigación, el
24 Protocolo Homologado de Búsqueda y demás disposiciones jurídicas aplicables. Artículo 41. Remisión de expedientes. 3333 VIERNES 5 DE MARZO DE 2021 / Número 26 ter.
25 ```
26
27 ### Fragmento 2 (similitud: 0.7015)
28
29 - **Archivo:** 03-05-21-ter ley general jalisco.pdf
30 - **Documento ID:** 7b4902c1-ed59-4ea6-b498-b2d04f120eac
31 - **Chunk:** 114
32
33 ```text
34 s relacionados con el ejercicio de las
35 funciones y atribuciones de la Comisión de Búsqueda; XXVIII. Proponer al Ministerio Público de la Federación a través de la Comisión Nacional, el
36 ejercicio de la facultad de atracción de conformidad con lo dispuesto en la Ley General; XXIX. Dar vista a las autoridades competentes en materia de responsabilidades
```

Figure 5. Content of the Markdown generated by Filtrum-Safety, including retrieved chunks, source file, chunk number, and similarity scores.

rag-context (5).md
Archivo

¿Cuáles son los criterios de interpretación que obliga fundamentar un amparo por omisión de búsqueda?

Si te refieres a un amparo indirecto contra la omisión de búsqueda de una persona desaparecida, la autoridad responsable no sólo está obligada a realizar actos de búsqueda, sino que debe hacerlo conforme a ciertos criterios constitucionales, convencionales y legales. Al impugnar la omisión, normalmente se fundamenta en los siguientes criterios de interpretación:

1. Principio pro persona

Previsto en el artículo 1º constitucional y en los tratados internacionales de derechos humanos.

Implica que, ante varias interpretaciones posibles, debe adoptarse aquella que otorgue la protección más amplia a la persona desaparecida y a sus familiares.

Aplicación en el amparo:

- La obligación de búsqueda debe interpretarse de manera amplia.
- Las facultades de las autoridades no pueden restringirse mediante formalismos.
- Toda duda debe resolverse favoreciendo la localización de la persona.

2. Interpretación conforme

También derivada del artículo 1º constitucional.

Las normas de la Ley General en Materia de Desaparición de Personas deben interpretarse de conformidad con:

- La Constitución.
- La Convención Americana sobre Derechos Humanos.
- La Convención Internacional para la Protección de Todas las Personas contra las Desapariciones Forzadas.

Aplicación:

La autoridad no puede alegar ausencia de denuncia, falta de competencia o insuficiencia de indicios para abstenerse de buscar.

Figure 6. LLM response using the Markdown as context.

D. Supplemental cross-model evidence

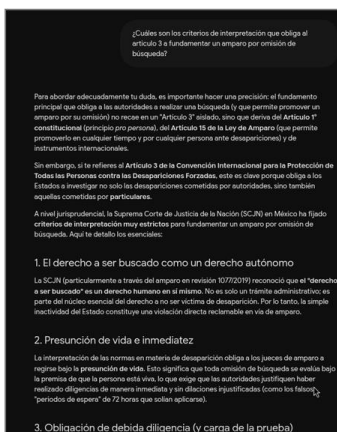
The following screenshots document the supplemental cross-model check. They are included as qualitative evidence only: they show how three different LLM interfaces responded to the same legal query with and without the Markdown context generated by Filtrum-Safety.

Direct prompting without Filtrum-Safety Markdown

Claude - direct response



Gemini - direct response



DeepSeek - direct response

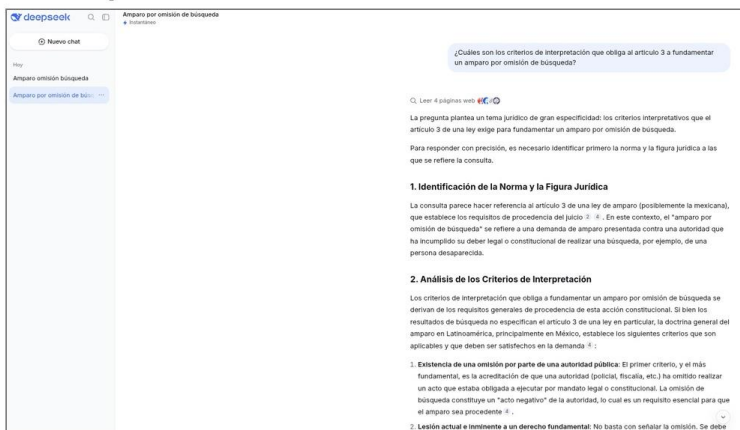
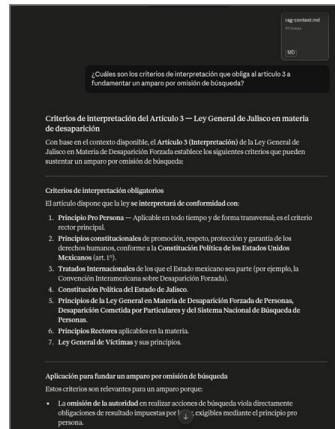


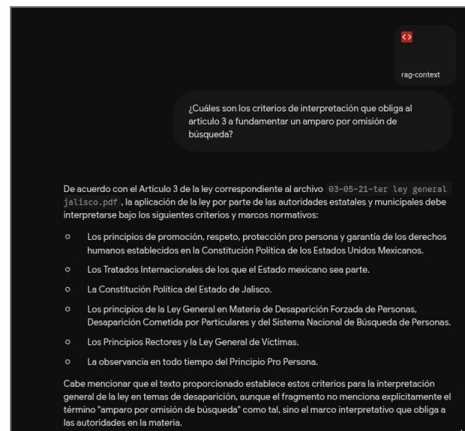
Figure 7. Direct outputs from Claude, Gemini, and DeepSeek without Filtrum-Safety Markdown context.

Prompting with Filtrum-Safety Markdown context

Claude - Markdown-assisted response



Gemini - Markdown-assisted response



DeepSeek - Markdown-assisted response

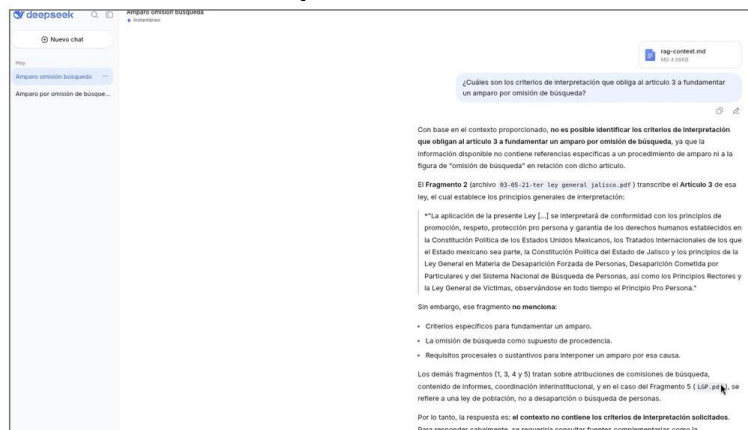
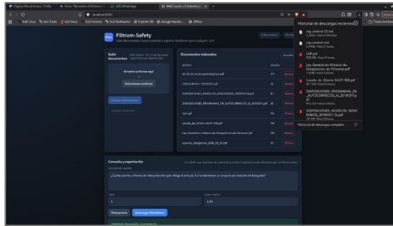


Figure 8. Markdown-assisted outputs from Claude, Gemini, and DeepSeek using Filtrum-Safety context.

Filtrum-Safety retrieval evidence for supplemental check

Filtrum-Safety interface



Retrieved fragments preview

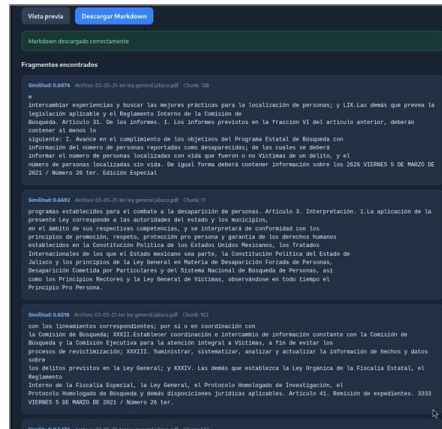


Figure 9. Filtrum-Safety retrieval interface and retrieved-fragment preview used to generate the supplemental Markdown context.

LLM Usage Statement

AI tools were used as support during the hackathon: Perplexity for initial literature exploration, Gemini for style correction, and Cursor to accelerate the construction of the preliminary prototype codebase. The team reviewed the final content, adjusted the technical claims, and avoided presenting observations from a preliminary validation as definitive conclusions. Claude, Gemini, and DeepSeek were also used as evaluation targets in the supplemental cross-model check; their outputs are treated as experimental evidence, not as independent legal authorities.

Any legal response generated with the system should be considered supporting material and not legal advice. The retrieved sources must be reviewed by a person before being used in academic, professional, or judicial activities.

Acknowledgements

We especially thank Daniel Hernandez Mota for his guidance and feedback during the development of the project. His comments contributed to improving the methodological clarity, scope, and final presentation of Filtrum-Safety.