PermitTEC v0.1: Standardized Metadata Corpus of NEPA Litigation Documents

Kaustav Bhattacharjee[1,+], Narmadha M Mohankumar[1,+], Julianna Puccio[1,+], Sayak Mukherjee[1], Laren Spear[1], Olivia Hess[1], Rizwan Ashraf[1], Thomas Serrano[1], Ellyn Ayton[1], James Bandy[1], Brian Chen[1], Renuka Chintalapati[1], Sam Donald[1], Milan Jain[1], Michael Kieburtz[1], Cole Man[1], Sai D. Koneru[1], Paul Rigor[1], Joshua Wassing[1], Gregory Wint[1], William Zhang[1], Dave Goodman[1], Jim Jackson[1], Beau Morton[1], Rounak Meyur[1], Anurag Acharya[1,*], Sai Munikoti[1,*], and Sameera Horawalavithana[1,*]

[1]Pacific Northwest National Laboratory
[+]These authors contributed equally to this work (ordered alphabetically)
[*]Corresponding authors: {anurag.acharya, sai.munikoti, yasanka.horawalavithana, permitai}@pnnl.gov

ABSTRACT

The National Environmental Policy Act of 1969, as amended (NEPA), mandates that federal agencies assess and document potential environmental impacts before deciding on proposed actions. While significant progress has been made in cataloging and standardizing NEPA documents themselves, the legal challenges that frequently arise from these decisions remain poorly cataloged and largely inaccessible for systematic analysis. Litigation challenging NEPA compliance can substantially delay project timelines, reshape agency decision-making, and establish precedents that influence future environmental reviews — yet no standardized, machine-readable corpus exists that links litigation records to the NEPA projects they contest. The absence of such a resource limits empirical understanding of litigation patterns, impedes risk assessment, and constrains evidence-based efforts to modernize the permitting process. In this work, we publicly release PermitTEC v0.1, a curated metadata corpus of 761 federal court litigation cases related to NEPA and adjacent environmental statutes. The corpus is constructed through an NLP pipeline that employs large language models (LLMs) for extracting contested project references from litigation text and few-shot classification to categorize each case by the nature of its legal challenge — whether it contests a specific NEPA document (e.g., an EIS, EA, or FONSI) or the absence of a required environmental review.To bridge the gap between litigation records and NEPA project data, we develop and evaluate three complementary matching approaches: LLM-based keyword extraction, fuzzy matching with composite metadata keys, and semantic retrieval, for linking litigation cases to their corresponding project records in NEPATEC v2.0, a corpus of over 140,000 NEPA documents spanning 60,000 projects across more than 60 federal agencies. Together, PermitTEC v0.1 and NEPATEC v2.0 form an integrated permitting-to-litigation data infrastructure that will enable downstream applications, including litigation trend analysis, project-level risk prediction, identification of recurrent grounds for legal challenge, precedent retrieval for legal practitioners, and AI-assisted compliance review — advancing the broader effort to modernize federal environmental permitting through data-driven insights. The PermitTEC v0.1 Dataset is publicly accessible at https://huggingface.co/datasets/PNNL/PermitTECv0.1.

1 Background & Summary

Federal agencies in the United States are required under the National Environmental Policy Act of 1969 (NEPA) to evaluate the potential environmental consequences of proposed actions and to document their findings before reaching a decision. The NEPA review process produces documents at varying levels of analytical depth: categorical exclusions (CEs) for actions that normally do not have significant environmental effects, environmental assessments (EAs) for actions whose significance is uncertain, and environmental impact statements (EISs) for major federal actions with potentially significant environmental impacts. These documents, together with associated records of decision (RODs) and findings of no significant impact (FONSIs), constitute the administrative record upon which agency decisions rest. The preparation, process, adequacy, and sometimes the

very existence of these documents are frequently subject to legal challenge. NEPA litigation shapes agency behavior, establishes legal precedent, and can substantially affect whether and when permitted projects proceed.

NEPA itself does not contain a private right of action or citizen suit provision. Instead, plaintiffs (including environmental organizations, state and local governments, tribal nations, industry groups, and affected individuals)typically bring NEPA challenges under the Administrative Procedure Act (APA), 5 U.S.C. § 706, which authorizes judicial review of final agency actions. Under the APA, courts review agency compliance with NEPA under the "arbitrary and capricious" standard (APA § 706(2)(A)), evaluating whether the agency took the requisite "hard look" at the potential environmental consequences of its proposed action and whether the resulting decision was reasonable and adequately supported by the administrative record.

NEPA challenges are generally filed in United States District Courts and may be appealed to the United States Courts of Appeals. The range of federal actions that give rise to NEPA litigation is broad. Common grounds for legal challenge include:

- **Failure to prepare a NEPA document** when one was required for a major federal action significantly affecting the environment;
- **Inadequacy of an EIS or EA**, such as failure to analyze a reasonable range of alternatives, insufficient consideration of direct, indirect, or cumulative impacts, or reliance on outdated or incomplete data;
- **Improper reliance on a categorical exclusion**, where a plaintiff argues that the action does not qualify for exclusion from detailed environmental review due to extraordinary circumstances or significant environmental effects;
- **Failure to supplement** an existing EIS or EA in light of significant new information or substantially changed circumstances; and
- **Challenges involving adjacent environmental statutes** that are often intertwined with NEPA compliance, such as the Endangered Species Act (ESA), the Clean Water Act (CWA), the National Historic Preservation Act (NHPA), or the Administrative Procedure Act (APA) itself.

Importantly, not all litigation in the environmental permitting space directly challenges a specific NEPA document. Some cases challenge the absence of a required environmental review — for example, arguing that an agency improperly proceeded without preparing an EA or EIS — while others primarily contest compliance with other statutes or requirements, with NEPA claims appearing as secondary or ancillary grounds. This heterogeneity in the nature of legal challenges necessitates a classification framework that distinguishes among these categories, as the implications for permitting practice differ substantially depending on what is being contested.

1.1 Dispositions in NEPA Litigation

The outcome of a NEPA litigation case, i.e., its disposition, can take a wide variety of forms depending on the court's findings regarding the merits of the NEPA claim, the procedural posture of the case, and the appropriateness of available remedies. Understanding the distribution and character of these dispositions is essential for analyzing how litigation affects project timelines, agency decision-making, and environmental protection. The principal disposition categories observed in NEPA litigation can be organized into three broad groups:

Merits-Based Dispositions: When a court reaches the substance of a NEPA challenge, several outcomes are possible. In a *remand without vacatur*, the court identifies a NEPA violation but permits the challenged agency action to remain in effect while the agency corrects the identified deficiency — a remedy frequently applied when vacating the action would cause significant practical disruption or when the deficiency appears curable without altering the ultimate decision. In a *remand with vacatur*, the court finds a NEPA violation and nullifies the agency's decision, returning the matter to the agency for preparation of adequate environmental review; this is the default remedy under APA § 706(2) and is imposed when the deficiency is serious, and the equities do not favor leaving the action in place. Courts may also issue an *injunction*, either permanent or preliminary, directing the agency or project proponent to halt or modify specific activities. Preliminary injunctions are evaluated under a four-factor test assessing the likelihood of success on the merits, the risk of irreparable harm in the absence of relief, the balance of equities between the parties, and whether the injunction would serve the public interest[1]. A *declaratory judgment for the plaintiff* establishes that the agency violated NEPA without necessarily ordering specific injunctive or vacatur relief, while a *judgment for the defendant* or *summary judgment for the federal defendant* concludes that the agency complied with NEPA's requirements and is entitled to judgment as a matter of law.

Procedural and Jurisdictional Dismissals: Many NEPA cases are resolved on threshold procedural or jurisdictional grounds before a court addresses the merits. A case may be *dismissed for lack of standing* when the plaintiff cannot demonstrate a concrete and particularized injury in fact that is fairly traceable to the challenged action and redressable by a favorable court

decision. Cases may also be *dismissed as not ripe* when the agency action is not yet sufficiently final, or the alleged harm remains speculative, or *dismissed as moot* when the underlying controversy is no longer live, such as when a challenged project has been completed, abandoned, or substantially modified. A court may *dismiss for failure to exhaust administrative remedies* when the plaintiff did not raise its NEPA concerns during the administrative process — for example, during the public comment period on a draft EIS — and the court determines that such participation was a prerequisite to judicial review. Dismissal may also occur for *lack of final agency action* under the APA (which generally limits judicial review to final agency actions), for l*ack of subject matter jurisdiction* (when the dispute falls outside the court's statutory authority), or *on statute of limitations grounds* (when the claim was filed outside the applicable limitations period).

Appellate Dispositions: When a district court decision is appealed, the appellate court may *affirm* the lower court's ruling, *reverse* it, or *reverse and remand* the case for further proceedings consistent with the appellate court's opinion. Appellate review in NEPA cases has played a particularly significant role in shaping the doctrine, as these decisions may establish precedent on issues such as the adequacy of alternatives analysis, the scope of cumulative impact review, and the appropriate remedy for identified NEPA violations.

1.2 The Data Gap in NEPA Litigation

Despite the central role that litigation plays in shaping NEPA implementation and federal permitting outcomes, there is no publicly available, standardized, machine-readable corpus that systematically catalogs NEPA litigation cases with structured metadata. Existing legal databases and repositories provide access to case text but do not offer NEPA-specific metadata fields such as the type of NEPA document challenged, the lead federal agency, the nature of the underlying environmental action, the disposition category, or critically, a link to the underlying NEPA project record in any permitting database. Court opinions reference contested projects using informal, abbreviated, or colloquial names that diverge substantially from the canonical project titles maintained in federal agency systems. For example, a court decision may refer to a project by a shortened name, a geographic description, or an agency-internal identifier that bears little resemblance to the standardized title recorded in a NEPA database.

This absence of structured litigation data limits the ability of researchers, policymakers, legal practitioners, and agencies to systematically analyze litigation patterns and outcomes. Without a standardized corpus, fundamental questions remain difficult to answer at scale: Which types of NEPA documents are most frequently challenged? Which agencies face the highest litigation rates, and for what categories of actions? How do disposition outcomes vary across jurisdictions, project types, and time periods? What procedural or substantive characteristics of NEPA documents are associated with greater litigation resilience? Answering these questions requires not only a curated collection of litigation records with structured metadata but also a reliable mechanism for linking those records to the NEPA project data they contest.

NEPATEC v2.0 established a foundational corpus of more than 140,000 NEPA documents from over 60,000 projects spanning more than 60 federal agencies, with standardized metadata aligned to the metadata standards recommended by the Council on Environmental Quality (CEQ)[2]. **Permit Te**xt **C**orpus (PermitTEC v0.1) extends this data infrastructure into the litigation domain, providing a curated metadata corpus of 761 federal court cases related to NEPA and adjacent environmental statutes. Each court decision is processed through a natural language processing (NLP) pipeline that extracts contested project references from unstructured judicial text, classifies the nature of the legal challenge — distinguishing whether the case contests a specific NEPA document, the absence of a required environmental review, or compliance with an adjacent statute — and maps each case to its corresponding NEPA project record in NEPATEC v2.0 through multiple complementary matching strategies.

Where NEPATEC v2.0 captures the administrative dimension of NEPA (what agencies prepared and decided), PermitTEC v0.1 captures the adjudicative dimension (how those decisions were challenged and what courts concluded). Together, these two corpora enable an integrated view of the federal environmental permitting lifecycle, from the initiation of environmental review through the resolution of legal disputes. This integrated infrastructure supports a range of analytical and applied tasks that neither corpus could support alone: tracing the complete trajectory of a federal action from environmental review through legal challenge to judicial resolution; constructing agency-specific and project-type-specific litigation risk profiles; analyzing temporal and geographic trends in NEPA litigation volume and outcomes; evaluating whether particular document preparation practices, levels of public engagement, or analytical approaches correlate with resilience to legal challenge; and developing predictive models that flag proposed actions with elevated litigation risk based on environmental, procedural, and contextual characteristics. By bridging the gap between permitting records and litigation outcomes, PermitTEC v0.1 contributes to the broader effort to modernize federal environmental permitting through empirical evidence, transparency, and data-driven decision-making.

1.3 The Challenges Associated with Linking Court Decisions to NEPA Documents

Linking court decisions to the underlying NEPA documentation is difficult and while it can be streamlined using an LLM, it typically requires human review and validation. Court opinions are not written in a uniform format across jurisdictions, and they often do not clearly identify the NEPA document being challenged. In some cases, the decision explicitly states the project name, the type of NEPA review (CE, EA, or EIS), the year, and other identifying details. More often, however, this information is missing, incomplete, or inconsistent. In some decisions, the court acknowledges that NEPA is at issue but does not identify the underlying NEPA document at all.

Additional challenges include:

- **No NEPA document to link:** Some lawsuits allege that a federal agency failed to prepare any NEPA document. These decisions are NEPA-related, but there is no CE/EA/EIS to associate with the case.
- **NEPA mentioned only in passing:** Some cases reference NEPA as an analogy or in the background section of the case, while the claims and outcome are based on a different statute or regulation.
- **Document not available:** In some instances, the relevant NEPA document can be identified, but it is not in NEPATEC and cannot be found online (a problem that is more common in older cases).
- **Procedural rulings:** Some cases are resolved on procedural grounds rather than the merits of NEPA compliance, making the underlying NEPA document less relevant to the disposition.
- **Amendments and tiered decisions:** Some cases challenge an amendment, framework, or other decision that tiers to or follows from an earlier EA or EIS. The parent document may be available, but the amendment often is not. These relationships can be difficult to distinguish reliably, particularly when multiple amendments stem from the same underlying NEPA document.
- **Post-litigation documents mistakenly matched:** Automated approaches (including LLMs) may identify NEPA documents prepared after the court decision—often in response to remand—as potential matches. This can usually be addressed by requiring that any matched NEPA document predate the court decision.
- **Challenges to multiple documents:** Some cases involve many NEPA documents, making matching difficult at scale. For example, 839 F.3d 1276 involved a challenge to BLM New Mexico's approval of 260 Applications for Permit to Drill, supported by 260 separate EAs.

2 PermitTEC v0.1 Dataset

We introduce PermitTEC v0.1, a standardized public metadata dataset for NEPA-related federal litigation. PermitTEC is designed to complement NEPATEC v2.0: whereas NEPATEC captures the administrative record of environmental review, PermitTEC captures the adjudicative dimension through structured metadata describing how those decisions were challenged in court and, when possible, which underlying NEPA projects those cases contest. The dataset is publicly available at https://huggingface.co/datasets/PNNL/PermitTECv0.1.

The current public release focuses on litigation metadata and includes 761 validated records corresponding to federal court cases related to NEPA and adjacent environmental statutes. Each record is organized around a case and includes structured fields such as case title, citation, court, circuit, parties, ruling date, and prevailing party, together with project-linkage fields indicating whether the case was matched to a NEPATEC v2.0 project and, if so, the associated project identifier and contested project name. To support transparency and downstream quality control, each metadata field carries a provenance tag indicating whether the value was accepted from automated extraction, manually corrected, or left without manual review.

A core feature of the release is its integration with NEPATEC v2.0. Of the 761 validated litigation records, 223 (29.3%) were linked to NEPATEC project records, corresponding to 168 unique NEPA projects. The remaining records either do not directly challenge a NEPA document or involve challenges whose underlying project could not yet be identified in the current NEPATEC release. This makes PermitTEC v0.1 useful both as a standalone litigation metadata resource and as part of a broader permitting-to-litigation data infrastructure.

At a high level, the current release is dominated by appellate decisions, spans litigation from the 1970s to 2025, and links primarily to Environmental Impact Statement projects. As a result, PermitTEC v0.1 provides a concise but policy-relevant view of the litigation dimension of environmental permitting and supports downstream analyses of litigation patterns, project-level risk, and precedent-informed permitting research.

Figures 1–4 illustrate some key highlights of the PermitTEC v0.1 dataset.

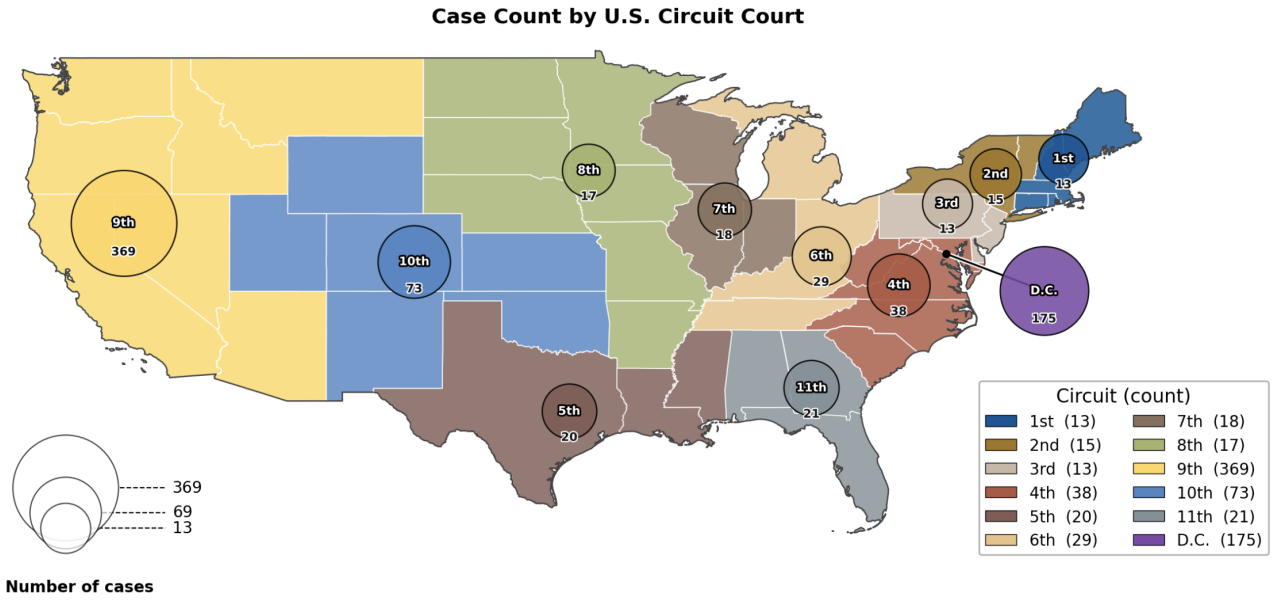


Figure 1. Geographic distribution of cases across U.S. Circuit Courts—states shaded by circuit, with proportional bubbles showing the number of cases in each circuit. Please note that the map looks distorted compared to a traditional Mercator projection map because an angular shape is being displayed in a flat surface.

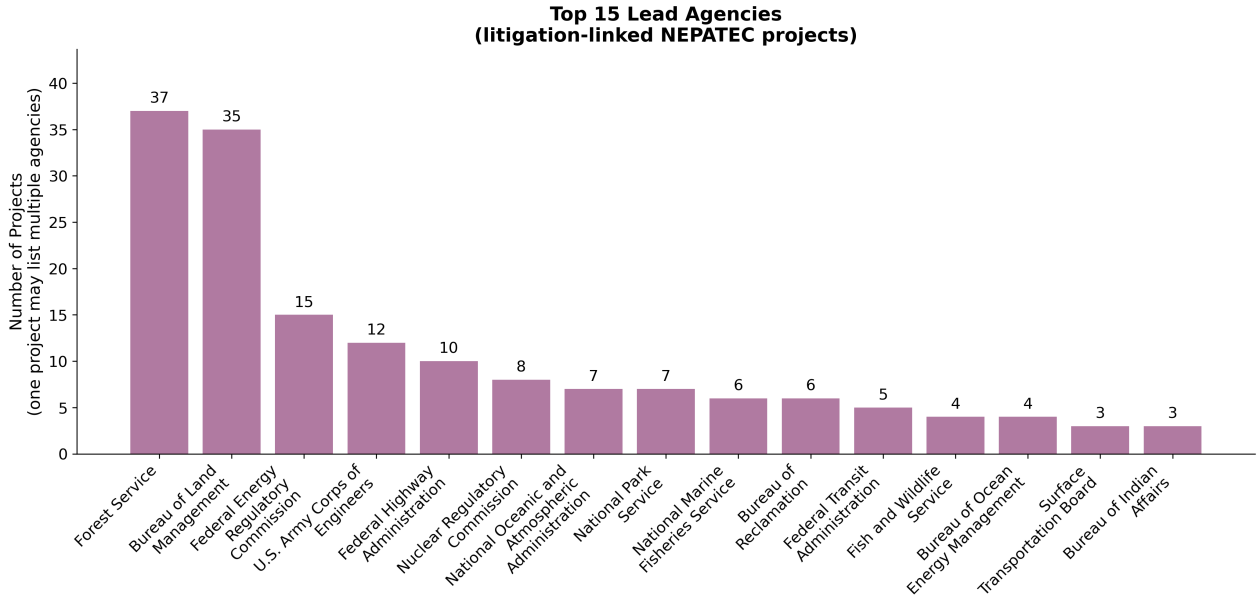
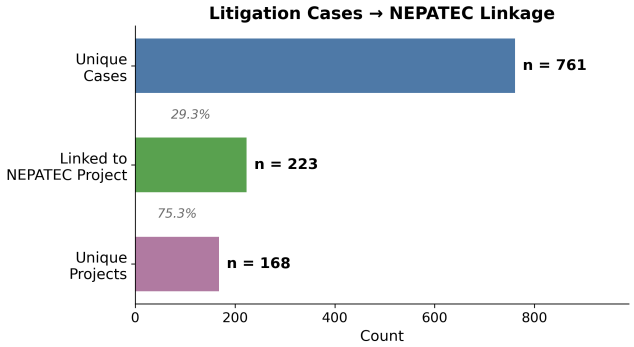
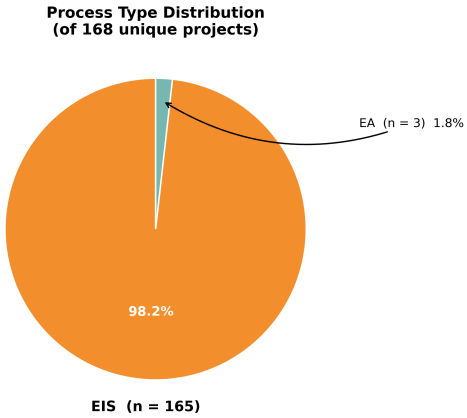


Figure 2. Distribution of the top 15 lead federal agencies among litigation-linked NEPA projects. Agency metadata was also retrieved from NEPATEC v2.0 project records by cross-referencing with PermitTEC. Note that a single project may list multiple lead agencies.



(a) Litigation-to-NEPATEC project linkage funnel. Of the 761 unique litigation cases in the corpus, 223 (29.3%) were successfully linked to a NEPA project record in NEPATEC v2.0, corresponding to 168 unique NEPA projects (75.3% of linked cases map to distinct projects). The linkage was established using the three complementary matching methods described in Section 3.3.



(b) Distribution of NEPA process types among the 168 unique litigation-linked projects. Environmental Impact Statements (EIS) account for 98.2% (n = 165) of linked projects, with Environmental Assessments (EA) comprising the remaining 1.8% (n = 3). Process type metadata was obtained from the corresponding NEPATEC v2.0 project records via the matched project UUID.

Figure 3. Litigation-to-project linkage outcomes and NEPA process type distribution for the subset of litigation-linked projects in PermitTEC v0.1.

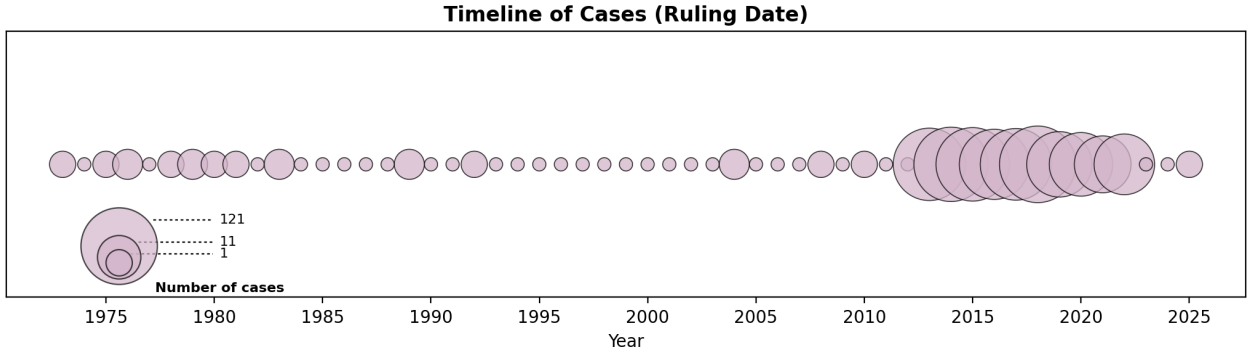


Figure 4. Timeline of Ruling Dates for litigation cases. Dates of ruling for litigation cases range from the 1970s to 2025, with the majority of cases occurring between 2010 and 2023

2.1 Dataset Structure

PermitTEC v0.1 is a metadata dataset organized around case-level records. Each record in PermitTEC represents a case (litigation or permitting action) with associated metadata, project linkages, and where available, page-level document text. Every field within `case_metadata` includes both a `value` and a `source` tag indicating data provenance. Table 1 describes the core metadata attributes. A representative PermitTEC v0.1 record is shown in Listing 1 in Appendix B. The example illustrates the case-centered schema, the organization of metadata into nested fields, and the explicit recording of provenance through per-field source tags.

3 Methods

3.1 Litigation Document Collection & Preprocessing

The construction of the PermitTEC v0.1 litigation corpus follows a three-stage process: identification of relevant litigation cases from an authoritative seed database, acquisition of full court opinion documents from legal repositories, and preprocessing of raw documents into clean, machine-readable text with structured metadata. This subsection describes each stage. Figure 5 illustrates the resulting hierarchical data model, in which each litigation case is represented as a top-level entity with a unique identifier (UUID), case-level metadata, and one or more associated court documents, each carrying its own document-level

Table 1. Metadata schema for the PermitTEC v0.1 litigation corpus. Case metadata fields characterize the litigation case itself, while NEPA project association fields capture the linkage to corresponding project records in NEPATEC v2.0.

Field Group	Metadata Field	Definition
Case Metadata	Case Title	Full official title of the legal case as it appears in court records (e.g., *Center for Biological Diversity v. Bureau of Land Management*).
	Case Citation	Formal legal citation for locating the case in a legal database, including reporter, volume, and page (e.g., 698 F.3d 1101, 9th Cir. 2012).
	Court	Name of the court where the case was heard (e.g., U.S. District Court, District of Nevada; Ninth Circuit Court of Appeals).
	Circuit	Federal judicial circuit or court jurisdiction in which the case was decided (e.g., 9th Circuit, D.C. Circuit, Federal Circuit).
	Plaintiff	Party or parties initiating the lawsuit.
	Defendant	Party or parties named in the lawsuit.
	Prevailing Party	The party that won the case (Agency or Challenger).
	Date of Ruling	Date on which the court issued its final ruling or decision (ISO 8601 format: YYYY-MM-DD).
NEPA Project Association	In NEPATEC	Boolean flag indicating whether this litigation case has been linked to a known NEPA project record in the NEPATEC database (True / False).
	NEPA Project Related Keywords in Litigation	Key NEPA-related terms, project names, document titles, or regulatory citations extracted from the litigation record that were used to match this case to a NEPATEC project.
	NEPATEC Project UUID	Universally unique identifier of the associated NEPATEC project record. Enables direct cross-reference to the NEPATEC metadata schema and document repository.
	Contested Project Name	Title of the matched NEPA project as recorded in the NEPATEC database.

metadata and paginated text content. Table 1 provides the complete definition of each metadata field in the corpus, organized into two groups: *case metadata* fields that characterize the litigation case itself, and *NEPA project association* fields that capture the linkage between the litigation case and its corresponding project record in NEPATEC v2.0. Together, the figure and table define the structural and semantic schema of the PermitTEC v0.1 corpus.

Source Data and Case Identification: The seed database of NEPA-related litigation cases was provided by the Breakthrough Institute, a non-partisan research organization that has conducted extensive analysis of environmental permitting and litigation trends[3]. This database contains case identifiers, basic citation information, and preliminary annotations for litigation cases involving NEPA and adjacent environmental statutes. After applying inclusion criteria to scope the corpus to federal court cases with a substantive connection to NEPA processes, we retained 761 cases as the foundation of PermitTEC v0.1.

Document Acquisition Using the case identifiers from the seed database, full court opinion documents were retrieved from two complementary legal repositories: Westlaw[4], a comprehensive commercial legal database providing broad coverage of federal court opinions with standardized citation metadata, and CourtListener[5], an open-access legal repository maintained by the Free Law Project that provides free access to federal court opinions. Documents were obtained in both PDF and HTML formats, depending on source availability. Where a case was available in both repositories, we applied deduplication based on case citation to avoid redundant records. Westlaw served as the primary source for coverage completeness, while CourtListener provided supplementary access for cases available in the public domain.

Document Preprocessing and Text Extraction: Raw litigation documents were processed through the text extraction layer of the automated ETL pipeline. For PDF documents, text extraction was performed using a PyMuPDF-based batch processor that parses document content on a page-by-page basis. HTML documents underwent tag stripping and structural normalization to isolate the judicial opinion text from surrounding navigation, header, and footer markup. Following initial extraction, all text underwent a multi-step normalization pipeline. Residual HTML entities were unescaped into their plaintext equivalents. Unicode NFKC normalization was applied to standardize typographic variants, ligatures, and compatibility characters into canonical form. Character encoding errors introduced during format conversion or OCR were detected and repaired through mojibake correction. Redundant whitespace, irregular line breaks, and extraneous formatting artifacts were regularized. Finally,

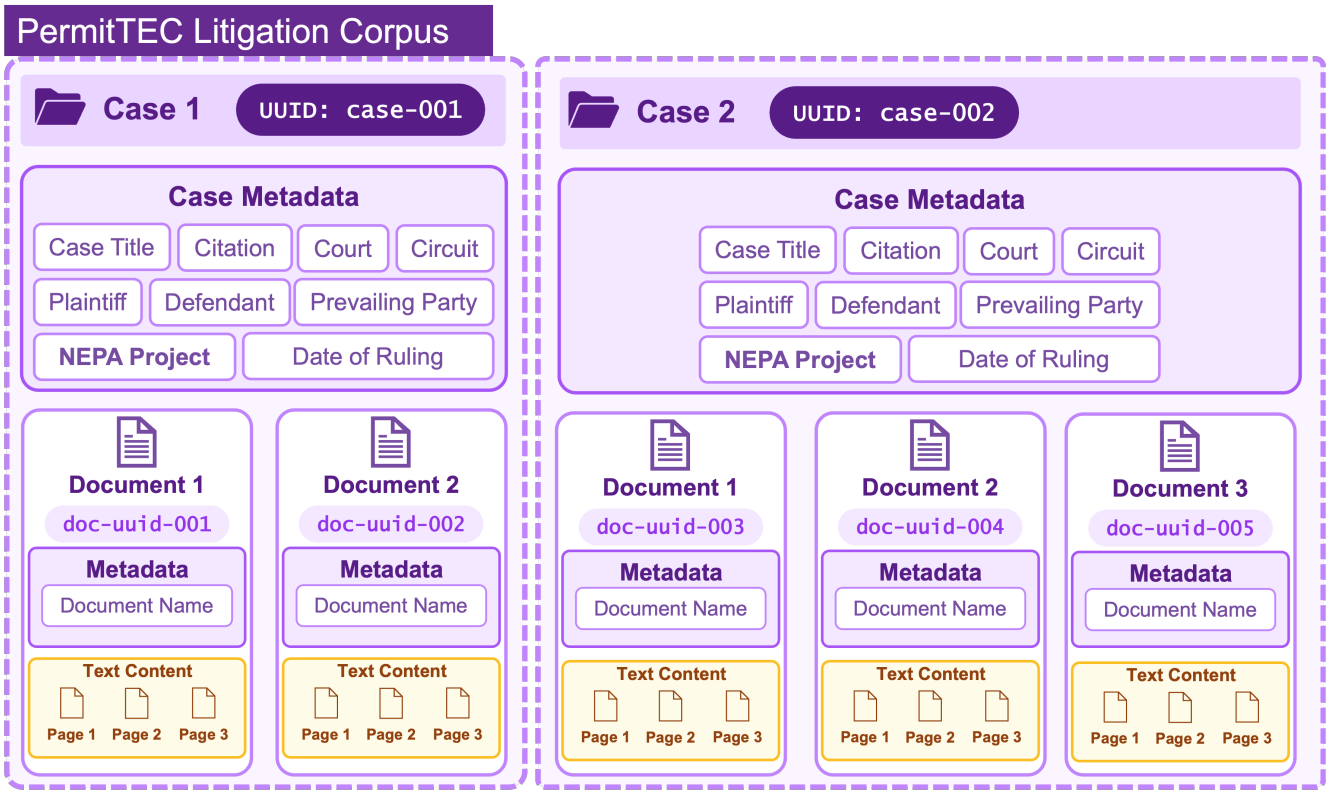


Figure 5. Hierarchical structure of the PermitTEC Litigation Corpus. Each case is assigned a unique identifier (UUID) and contains case-level metadata (case title, citation, court, circuit, plaintiff, defendant, prevailing party, associated NEPA project, and date of ruling) and one or more associated court documents, each with document-level metadata (document type, document title, and filing date). Paginated text content (shown in yellow) represents planned additions in future releases of the corpus; the current v0.1 release includes case-level and document-level metadata only.

case-insensitive deduplication was performed to identify and remove duplicate documents that may have been ingested from multiple sources or format variants. For each processed document, the pipeline computes quality and confidence metrics that characterize the reliability of extracted text for downstream processing. These metrics capture factors such as character-level extraction confidence, structural completeness (e.g., whether page boundaries are preserved), and the presence of OCR artifacts or encoding anomalies. Extracted text, computed metrics, and associated metadata are persisted into processed storage buckets on AWS S3, with relational metadata written to PostgreSQL following the PermitTEC schema. OpenSearch indexes both text and metadata to support full-text search and analytics across the corpus.

The resulting corpus consists of 761 litigation cases encompassing multiple court documents per case — including district, appellate, and supreme court decisions and related orders — organized in the hierarchical structure shown in Figure 5. Each case carries standardized case-level metadata (case title, citation, court, circuit, plaintiff, defendant, prevailing party, associated NEPA project, and date of ruling), while each document within a case carries document-level metadata (document type, document title, and filing date). Paginated text content for each document is planned for inclusion in future corpus releases. This two-level metadata architecture enables case-level litigation analysis within a single unified corpus, with document-level textual analysis to be supported upon the inclusion of paginated text content in subsequent releases. The automated extraction of these metadata fields from unstructured judicial text — and the methods used to populate, validate, and structure them at both the case and document levels — is described in the following subsections.

3.2 Case Metadata Extraction

3.2.1 Extraction Methodology

To systematically extract the case level metadata elements from unstructured text, we implemented an extraction pipeline using DSPy framework as a primary orchestration layer[6]. After the judicial documents were pre-processed into text, each document was passed through a series of metadata specific DSPy extractors with each extractor acting as a separate program operating over the same input text extracting one target field in the schema. This decomposition of the task into individual

Table 2. Case-level metadata fields extracted by the PermitTEC pipeline, with corresponding schema keys and evaluation metrics used to assess extraction quality against ground truth.

Metadata Field	Schema Key	Evaluation Metric
Case Title	`case_title`	Semantic Similarity (Embedding)
Citation	`citation`	Semantic Similarity (Embedding)
Court	`court`	Semantic Similarity (Fuzzy)
Circuit	`circuit`	Exact Match
Plaintiff	`plaintiff`	Semantic Similarity (Embedding)
Defendant	`defendant`	Semantic Similarity (Embedding)
Ruling Date	`ruling_date`	Exact Match
Prevailing Party	`prevailing_party`	Semantic Similarity (Embedding)
Disposition	`disposition`	Semantic Similarity (Embedding)

programs allows targeted prompt refinements. Furthermore, each extractor defines a default fallback output in the event of extraction failures when running on heterogeneous documents and partial extraction errors. This provides a modular interface to benchmark multiple model configurations while keeping the task definitions and the prompts fixed improving reproducibility of experiments and field level error analysis. The implementation also supports scalable batch processing for a given DSPy execution program enabling the pipeline to process large collections of litigation documents.

3.2.2 Benchmarking LLM Approaches for Case Metadata Extraction

Having established the prompt-based extraction methodology, we now evaluate its performance across the nine case-level metadata fields targeted by the pipeline. Table 2 lists each field, its corresponding schema key, and the evaluation metric used to assess extraction quality against ground truth. Evaluation metrics were selected based on the nature of each field: exact match for fields with constrained, unambiguous values (e.g., circuit designation, ruling date), semantic similarity using sentence embeddings for free-text fields, where surface-form variation is expected (e.g., casetitle, plaintiff, defendant), and fuzzy string similarity for fields with semi-structured but variable formatting (e.g., court name).

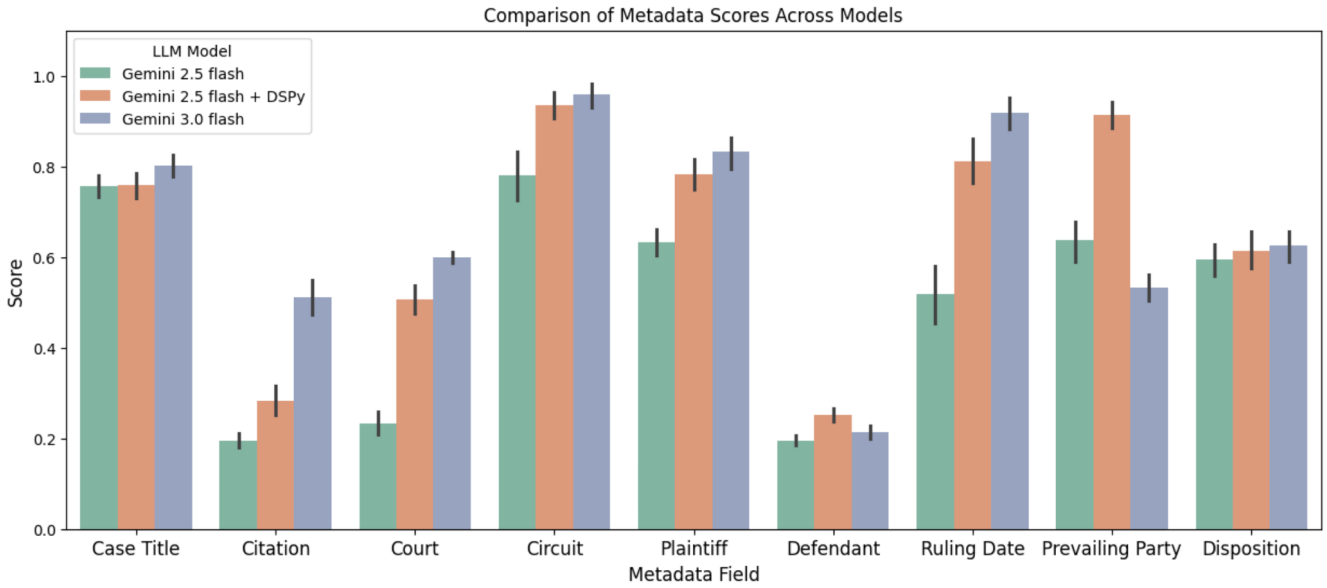
Dataset Description. The full corpus contains 856 litigation PDFs. Of these, 88 were excluded because they were unpublished, leaving 768 documents available for extraction. Among these, our pipeline failed to extract any usable text from 7 PDFs, resulting in a working dataset of 761 text-bearing documents used for downstream extraction tasks.

Benchmark Construction. To evaluate extraction performance, we constructed a benchmark dataset by aligning the litigation PDFs in the corpus with corresponding entries in the Breakthrough Institute (BTI) dataset, which provides pre-extracted metadata that serves as the ground truth. We used the complete set of 856 PDFs, as publication status did not affect their suitability for evaluation. Alignment was performed using case citation numbers, which were assigned as PDF file names during document acquisition and used to map each file to its corresponding row in the BTI dataset, yielding 698 matched cases. Because benchmarking required complete ground-truth labels across all nine metadata fields, we further restricted the benchmark to the 329 BTI-aligned cases containing non-null values for every field. Notably, all 329 cases in the resulting benchmark are district-level opinions, as appellate-level cases in the BTI dataset contained at least one null value among the nine target fields. This final benchmark set of 329 cases was used for all evaluation experiments.

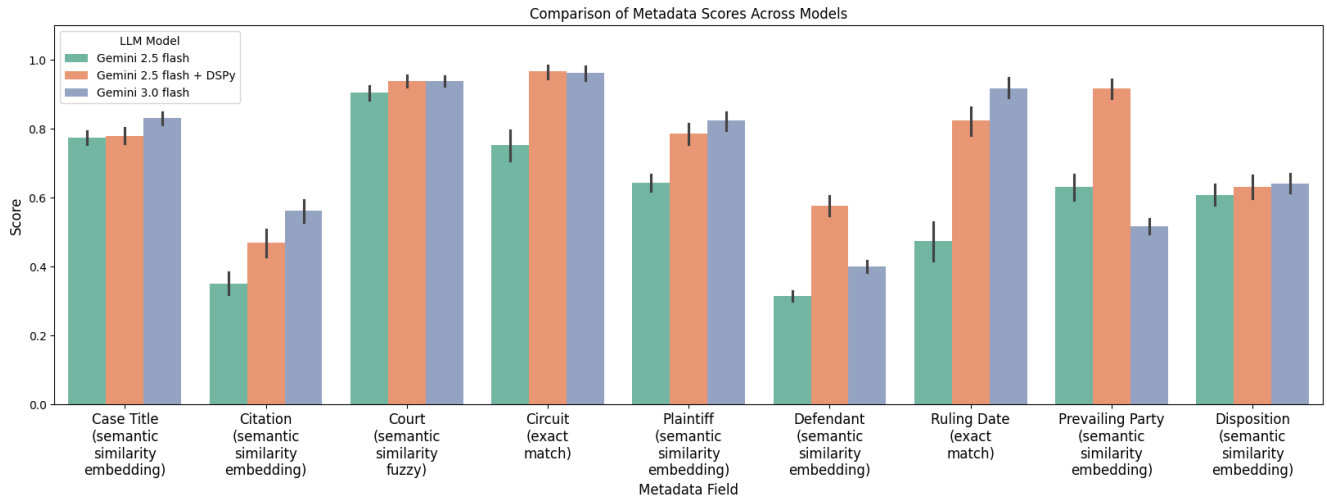
Model Comparison. We evaluated three model configurations for metadata extraction on the 329-case benchmark: Gemini 2.5 Flash, Gemini 2.5 Flash with DSPy-based prompt management, and Gemini 3.0 Flash (preview) Figure 6a presents the extraction scores across all nine metadata fields for each model configuration under the initial prompt design.

Prompt Optimization. Analysis of the initial results revealed that the Court and Circuit prompts were the primary candidates for improvement. These two fields are closely related — circuit designation is directly determined by the court in which a case is filed — and the initial prompts had been developed primarily using appellate-level cases as references, resulting in prompts that were overfit to appellate naming conventions and underperformed on district-level cases, which comprise the entirety of the benchmark set. Although DSPy offers automated prompt optimization capabilities, we opted for targeted manual prompt refinement for these two fields to maintain interpretability and control over the optimization process.

Figure 6b presents the extraction scores following prompt optimization. Court and Circuit both showed meaningful performance gains after refinement, confirming that the initial underperformance was attributable to prompt design rather than model capability. No additional prompt optimization was conducted for the remaining fields in the current release; further refinement across all metadata fields is planned for future versions of the corpus.



(a) Before prompt optimization.



(b) After prompt optimization.

Figure 6. Metadata extraction scores across nine case-level fields for three model configurations. Panel (a) shows results before prompt optimization. Panel (b) shows results after manual prompt optimization for the Court and Circuit fields; all other prompts remain unchanged. Evaluation metrics for each field are as specified in Table 2.

The nine metadata fields described above capture the *case-level* attributes of each litigation record — identifying who sued whom, in which court, and with what outcome. However, a critical piece of information for connecting litigation cases to the underlying NEPA projects they contest is not captured by these structured fields: the identity of the contested project itself. Court opinions reference projects using informal names, geographic descriptions, or agency-internal identifiers that bear little resemblance to the canonical project titles maintained in federal agency databases. Extracting these contested project references from unstructured judicial text requires a distinct approach, which we describe in the following subsection.

3.2.3 Extracting Project Titles/Excerpts from Litigation Documents

A critical prerequisite for linking litigation cases to their corresponding NEPA project records in NEPATEC v2.0 is identifying which project each case contests. Court opinions do not reference projects using standardized titles maintained in federal agency databases; instead, they describe contested actions through informal names, geographic descriptions, or narrative references embedded in the factual background of the opinion. To extract this information, we designed a prompt that identifies and

Table 3. State extraction logic by court level. District-level cases yield a single state, while appellate-level cases may yield multiple states depending on whether the originating jurisdiction can be identified from the opinion text.

Court Level	Extraction Logic	Output Format
District Court	State is identified directly from the district court name, which typically encodes the jurisdiction (e.g., *District of Oregon*, *Southern District of New York*).	Single state abbreviation (e.g., `OR`, `NY`)
Court of Appeals	Each circuit court spans multiple states. If the originating jurisdiction is explicitly stated in the opinion text, that state is extracted; otherwise, all states within the circuit are listed.	One or more state abbreviations, semicolon-delimited (e.g., `WA;OR;CA`)

extracts a short excerpt — a *contested project name excerpt* of up to four sentences — from each litigation document. The excerpt captures the passage in which the court describes the federal action giving rise to the dispute, typically found in the Background, Facts, or Introduction sections of the opinion. The full prompt specification, including extraction instructions and formatting constraints, is provided in Appendix C.1.

The prompt was applied to all litigation cases in the corpus using sequential extraction with Google Gemini 3 Flash. Preliminary efforts to map the extracted excerpts to NEPA project records in NEPATEC v2.0 were conducted on an initial subset ($n = 10$) using spaCy-based named entity recognition (NER) to identify project names within the excerpts, followed by fuzzy matching against NEPATEC v2.0 project titles. These early attempts were not consistently successful — even when NER correctly identified a project name, fuzzy matching frequently failed to return the corresponding NEPATEC2.0 record due to divergence between judicial language and canonical project titles. These findings motivated the development of more robust matching strategies, which are described in the following section.

A related observation during the extraction process was that not all litigation cases in the corpus directly challenge a specific NEPA document. Some cases contest the *absence* of a required environmental review, while others primarily challenge compliance with statutes adjacent to NEPA, such as the Clean Water Act (CWA) or the Endangered Species Act (ESA). This heterogeneity in the nature of legal challenges has implications for how extracted excerpts are interpreted and matched.

3.2.4 Geographic State Extraction from Litigation Documents

Beyond the contested project name excerpt, a second geographic signal was extracted to support the downstream task of mapping litigation cases to NEPA project records in NEPATEC v2.0: the U.S. state or states associated with each case. State-level geographic attribution serves as a complementary matching dimension — when combined with project name excerpts and agency information, it narrows the candidate space of potential NEPATEC v2.0 records and improves matching precision, particularly for cases where the contested project name alone is insufficient to identify a unique project. The full prompt specification is provided in Appendix C.1.

State extraction follows different logic depending on the court level at which the case was decided, as summarized in Table 3.

For district-level cases, extraction is straightforward as the state is directly derivable from the court name and a single abbreviation is returned. For appellate-level cases, the extraction is inherently less precise: each Court of Appeals encompasses multiple states, and the originating jurisdiction is not always stated in the opinion. In such cases, the prompt first attempts to identify the specific state of origin from contextual cues in the introduction or factual background of the opinion; if no state can be confidently identified, all states within the relevant circuit are returned as candidate jurisdictions. The extracted state information was integrated into the metadata extraction pipeline and used as a geographic signal in the litigation-to-project mapping methodology described in the subsequent sections.

3.3 Mapping Litigation Cases to NEPA projects

The preceding subsections established methods for extracting two key signals from each litigation case: a contested project name excerpt identifying the federal action under dispute, and the geographic state or states associated with the case. However, these extracted signals alone are insufficient for directly retrieving corresponding project records from NEPATEC v2.0. Litigation case documents frequently reference contested projects in ways that are noisy, informal, or incomplete — using shortened names, partial descriptions, or agency-internal identifiers that do not correspond cleanly to the standardized metadata within NEPA repositories. Because of these inconsistencies, the project names extracted from litigation text can diverge significantly from the canonical titles maintained in NEPATEC v2.0. This gap necessitates a matching process capable of tolerating variation in wording, accommodating differences in agency naming conventions and NEPA process descriptions, and leveraging geographic and administrative signals to disambiguate among the more than 60,000 project records in the corpus.

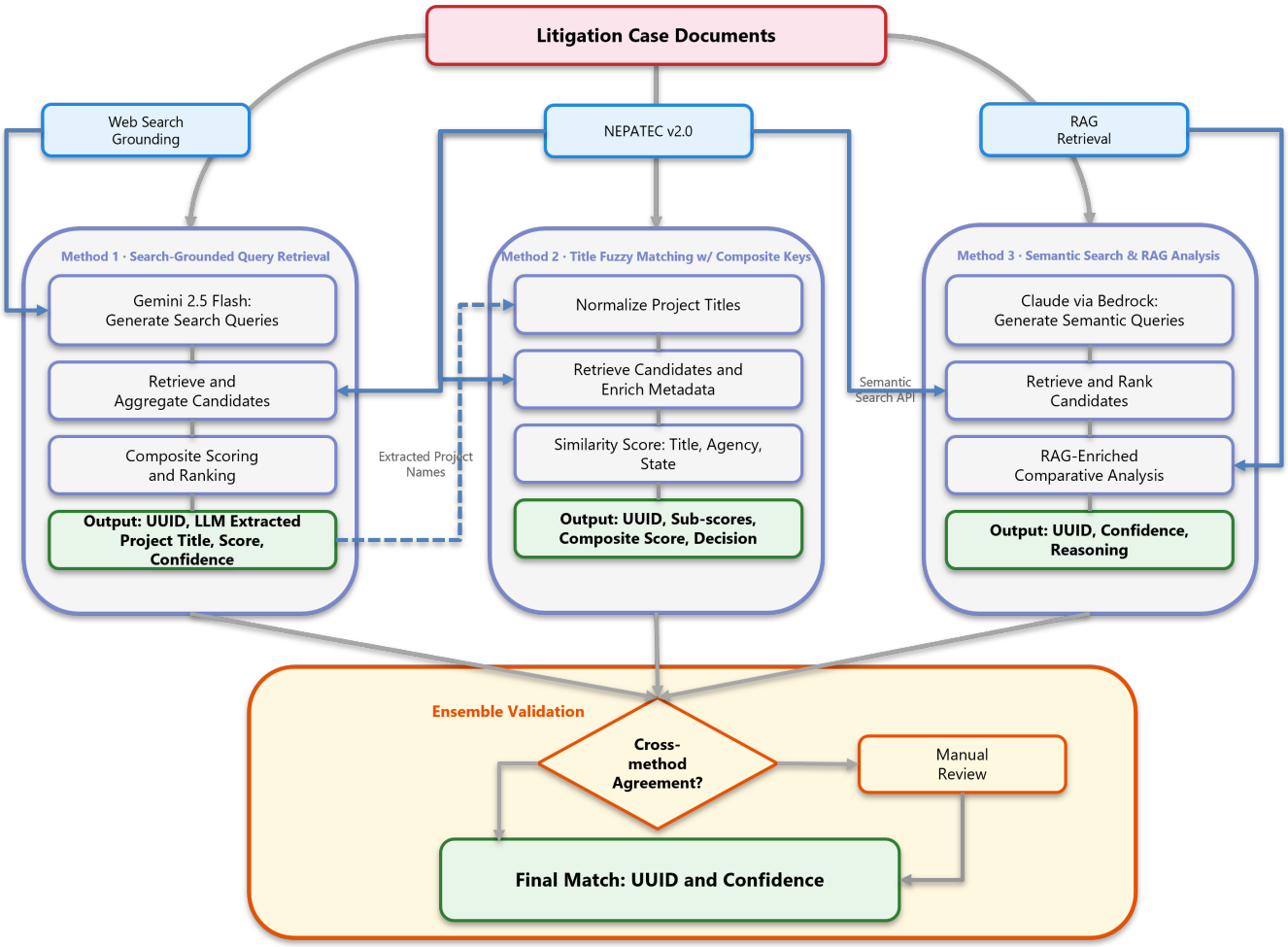


Figure 7. The entire litigation cases to NEPA project matching workflow via ensemble of the three matching methods. Blue arrows indicate data flow between modules, while grey arrows indicate logical flow of the process.

To address these challenges, we developed three complementary approaches for automatically linking litigation cases to their corresponding NEPA project records, each employing a distinct retrieval and ranking strategy. Method 1 uses search grounding through the Gemini Interactions API to extract project titles and structured query keywords directly from the litigation document, which are then used to query NEPATEC v2.0 project records. Method 2 takes the extracted project titles and applies fuzzy string matching combined with composite metadata keys — such as NEPA document type and state — to retrieve and score candidate projects based on multiple aligned signals. Method 3 employs an LLM-driven agent that generates targeted semantic search queries from the litigation text, retrieves candidate projects through semantic similarity, and applies retrieval-augmented generation (RAG) to reason over enriched project summaries before selecting the best match. The three methods are applied independently to evaluate how well each supports accurate and complete matching. These methods are shown in the context of the entire NEPA linking process in Figure 7, and each method is described in detail in the following subsubsections.

3.3.1 Method 1: Search-Grounded Information Extraction and Query-Based Retrieval

The first matching method employs an LLM-driven pipeline that extracts project-identifying information directly from litigation documents and uses it to retrieve candidate NEPA project records from NEPATEC v2.0. A central design principle underlying this method is the recognition that NEPA documents are prepared during the project planning and approval process — before litigation occurs — and therefore contain no references to case names, legal citations, or parties involved in subsequent legal challenges. Accordingly, the extraction stage is explicitly designed to look past the legal framing of the court opinion and focus on the characteristics of the *underlying environmental project*: its name, geographic location, responsible federal agency, project type, and associated environmental concerns. The method operates as a three-stage pipeline: query generation from litigation text, candidate retrieval against the NEPATEC v2.0 search index, and composite scoring to rank and select the best-matching project.

Stage 1: LLM-Based Query Generation. Rather than relying on pre-extracted text from the preprocessing pipeline, this method leverages a multimodal large language model — Gemini 2.5 Flash, accessed through the Google Interactions API — that processes litigation PDF files natively, without intermediate text extraction[7]. The model is augmented with search grounding capabilities, enabling it to verify and contextualize NEPA project references against live web sources during extraction. Given a litigation PDF, the model is prompted to generate the potential project title and a set of 5- 10 targeted search queries designed to retrieve the contested NEPA project from a structured environmental document repository. The prompt engineering strategy explicitly instructs the model to extract project-identifying characteristics rather than legal terminology. Queries must not contain case names, plaintiff or defendant names, legal citations, or court references, as none of these would appear in the target NEPA project records. The full prompt specification is provided in Appendix C.2.

Stage 2: Candidate Retrieval. The generated queries are executed against an OpenSearch index containing the full NEPATEC v2.0 corpus of project records. Each query is submitted independently, retrieving the top N matching project records ranked by OpenSearch relevance score. Because different queries emphasize different facets of the contested project — geographic, thematic, agency-based — they may each surface overlapping but non-identical candidate sets. Results from all queries are aggregated at the project level using the unique project UUID as the grouping key. For each candidate project, the aggregation captures the number of queries that returned it, the number of unique queries matched, the average relevance score across queries, and the maximum single-query relevance score.

Stage 3: Composite Scoring and Ranking. Aggregated candidate projects are ranked using a composite score that integrates multiple retrieval signals. The composite score accounts for both the breadth of query coverage — how many distinct queries surfaced a given project — and the depth of relevance — how strongly the project matched individual queries. Projects that are retrieved by a greater number of diverse queries and that achieve higher individual relevance scores receive correspondingly higher composite scores. A filtering threshold is applied to remove low-confidence candidates, and the highest-ranking project is returned as the predicted match along with a confidence assessment. The output for each litigation case comprises the best-match project UUID, the corresponding NEPATEC2.0 project title, the composite score, the confidence level, and the full ranked list of candidate projects to support manual review where needed.

Despite its ability to generate contextually rich queries grounded in web-verified project information, this method is sensitive to the quality and specificity of the LLM-generated queries. When the litigation document provides limited or ambiguous project details, the generated queries may be too broad to discriminate among candidate projects, leading to low-confidence matches. These limitations motivated the development of Method 2, which introduces a structured matching pipeline that leverages normalized metadata fields rather than free-text queries. A flowchart showing the overall method is shown in Figure 14 in Appendix E.

3.3.2 Method 2: Title Fuzzy Matching with Composite Keys

Method 2 addresses the variability of litigation-derived project references by transforming heterogeneous titles into normalized forms, retrieving candidate NEPA projects, enriching those candidates with document-level metadata, and computing a composite similarity score integrating title, agency, and state signals. The output is a ranked list of candidate projects, each accompanied by interpretable sub-scores and a final match decision based on a configurable threshold.

Stage 1: Title Normalization and Candidate Retrieval. The pipeline begins with a corpus of litigation case documents and their LLM extracted contested project names from Method 1. These titles are normalized using HTML unescaping, Unicode normalization (NFKC), correction of mojibake artifacts, whitespace cleanup, and case-insensitive deduplication. When a field contains multiple title variants—for example, separated by ";;"—each is independently cleaned and retained as part of a canonical title set suitable for fuzzy comparison. The normalized titles are then used to query OpenSearch for NEPA projects whose titles exhibit textual similarity. For each retrieved candidate, the system collects all associated NEPA document UUIDs.

Stage 2: Metadata Aggregation. The collected all associated NEPA document UUIDs are used to query the NEPATEC RDS service to obtain enriched metadata describing states, lead agencies, and document types. Because NEPA projects often span multiple documents—and therefore multiple states—the RDS response is aggregated at the project level to generate a comprehensive, deduplicated summary of each project's geographic extent and administrative attributes. Metadata normalization is essential for reliable comparison. Agency values are standardized through abbreviation mapping and hierarchical agency closure, enabling recognition of parent–child agency relationships (e.g., mapping "BLM" to include "DOI"). State normalization extracts canonical two-letter codes from free text, supporting U.S. states, territories, and Washington, D.C., and representing multi-state projects as comma-separated lists (e.g., "CA,NV,AZ").

Stage 3: Similarity Scoring. Using these normalized features, Method 2 computes three similarity signals: title similarity, agency compatibility, and state overlap.

- **Title Similarity**: Title similarity is the primary discriminative signal. Fuzzy matching techniques are applied, including token-set ratio to handle reordering or superstring relationships, and partial-ratio to capture substring matches typical of long NEPA titles. These fuzzy scores are combined with OpenSearch's retrieval score to generate a final title similarity value between 0 and 100.
- **Agency Compatibility**: Agency compatibility assesses whether the agencies mentioned in litigation align with those associated with the candidate NEPA project. The system first checks for intersections among standardized abbreviations (including parent tokens). Intersection yields a perfect score of 100. When abbreviation-level matching fails, the system evaluates token-based similarity across normalized agency names and abbreviations. Missing agency metadata returns "NA," which is handled neutrally during aggregation.
- **State Overlap**: State overlap evaluates geographic consistency. If at least one state code overlaps between the litigation-derived record and the NEPA project, the score is 100; absence of overlap yields 0. When either side lacks state metadata, the score is "NA." This ensures that missing data does not artificially penalize a candidate.

Stage 4: Composite Match Decision. The final similarity score is computed as a weighted combination of the three signals, with default weights of 0.5 for title, 0.25 for agency, and 0.25 for state. NA values are excluded during aggregation, preventing incomplete metadata from suppressing strong matches. A project is flagged as a match if its overall score exceeds a threshold (typically 80), balancing precision and tolerance for natural textual variation. A flowchart showing the overall method is shown in Figure 15 in Appendix E.

3.3.3 Method 3: LLM-Generated Semantic Search Queries and RAG Refinement

Method 3 employs an agent-based workflow that combines semantic retrieval with LLM-driven reasoning to map litigation case documents to NEPA projects. Rather than relying solely on title keywords or metadata fuzzy matching, the method treats the full litigation text as a queryable evidence source. It generates targeted search queries, retrieves broad candidate sets via semantic search, aggregates and ranks the results, and then applies retrieval-augmented analysis to select the most plausible project.

The Mapping Agent uses AWS Bedrock to call Claude for two tasks: (i) extracting targeted search queries from litigation text and (ii) conducting comparative analysis between the litigation case and NEPA project content. The agent returns a structured decision comprising the best-match project UUID, a confidence level, and detailed reasoning. Candidate discovery is performed via a semantic search API, and top candidates are enriched with project-level RAG summaries to support the final comparison step.

Stage 1: Query Generation (Claude via Bedrock). The litigation text is sent to Claude with instructions to produce 5–7 search queries focused on project title, location (state/city/county/geography), federal agency, project type, and environmental concerns.

Stage 2: Semantic Candidate Retrieval and Ranking. Each query is submitted to a semantic search API. For each query, the top 10–20 NEPA project results (by similarity score) are collected. Candidates from all queries are pooled and summarized by project. The agent maintains, per project, the title, agency, state, appearance count, and cumulative similarity score. Projects are ranked by the tuple `(appearances,total_score)` in descending order, and the top ten are retained for analysis. This cross-query aggregation prioritizes projects consistently retrieved under alternate phrasings and facets (title, agency, location, project type).

Stage 3: Retrieval-Augmented Generation (RAG)-enhanced comparative analysis. The top three candidates are enriched via RAG queries filtered by project UUID to retrieve detailed project information (e.g., purpose, location, lead agency, impacts). The agent presents (i) a summarized litigation case and (ii) detailed NEPA descriptions to Claude, prompting a structured comparison on scope, agency, location, environmental impacts, and timeline. Claude returns a best-match UUID, a confidence label (`high/medium/low`), and a narrative justification, along with alternative candidates if appropriate.

Decision Logic and Confidence. Claude's comparative analysis produces a structured decision comprising:

1. **Best-match**: the project most consistent with the litigation case.
2. **Confidence level**: one of `high`, `medium`, or `low`, indicating strength of alignment.
3. **Reasoning**: a textual explanation referencing scope, agency, geography, impacts, and timeline.
4. **Alternative matches**: Other plausible candidates.

A flowchart showing the overall method is shown in Figure 16 in Appendix E.

3.3.4 Comparative Development of the Three Matching Methods

The three methods were developed in sequence, each motivated by the limitations encountered in the previous approach.

Method 1 relied on search-grounded LLM extraction via the Gemini Interactions API to generate targeted queries from litigation PDFs and retrieve candidate projects from the NEPATEC v2.0 OpenSearch index. While effective for cases with explicit project descriptions, the method produced overly broad queries when project details were sparse (like redacted litigation documents), and its reliance on free-text retrieval signals alone — without structured metadata comparison such as agency alignment or state overlap — made it difficult to diagnose matching failures or systematically improve results.

These issues led to the development of Method 2, which introduced a structured fuzzy-matching pipeline using normalized titles, agency information, and state metadata. This method improved interpretability and offered more deterministic scoring. However, its accuracy was highly sensitive to metadata completeness. Many litigation records lacked explicit state references, and NEPA metadata frequently contained partial or inconsistent agency labels. Missing state fields and incomplete agency information made it difficult for Method 2 to compute reliable similarity scores, resulting in false negatives even when the correct project appeared among the candidates.

To address these shortcomings, Method 3 expanded the matching framework to operate directly on the full litigation text using semantic search and RAG-enhanced LLM comparison. This approach no longer depended on explicit keyword extraction or complete metadata fields. Instead, it identified projects through contextual similarity and then used detailed NEPA project descriptions to perform a grounded comparison. Method 3 proved particularly effective in cases where project names had changed, where titles were never explicitly stated in litigation, or where agency/state metadata were missing or incorrect.

In combination, the three methods form a progressively more robust pipeline. Method 1 provides broad initial recall through keyword extraction and title matching; Method 2 adds structured, interpretable scoring; and Method 3 handles the most challenging cases using semantic and contextual reasoning. Final validation compared agreement across methods and incorporated targeted manual review to confirm correct mappings.

3.4 Automated ETL Pipeline for Litigation Document Ingestion and Processing

To operationalize the PermitTEC v0.1 workflow as a reproducible and scalable system, we developed an automated extract-transform-load (ETL) pipeline for litigation document ingestion, text extraction, metadata enrichment, and integration into a searchable data lakehouse following the PermitTEC schema. The pipeline connects the previously described components — document preprocessing and text extraction, LLM-based case metadata extraction, contested project and state identification, and case-to-project matching — into a single end-to-end processing workflow.

The system uses a hybrid multi-cloud architecture that separates data authority and document lifecycle management from compute-intensive AI/ML inference. AWS serves as the authoritative environment for document ingestion, text extraction, relational storage, search indexing, and provenance tracking, while GCP provides the model-serving layer for large-scale metadata extraction and quality-controlled document intelligence. This design allows the storage and orchestration layer to scale independently from model-driven processing while preserving end-to-end traceability.

Although the current release focuses on NEPA litigation documents, the architecture is designed to be document-type agnostic. The same pipeline can be extended to additional federal and state environmental permitting corpora by adapting document-type-specific extraction prompts, metadata schemas, and validation rules without changing the core ingestion, orchestration, and integration framework. Additional implementation details, including the cross-cloud architecture and provenance model, are provided in Appendix A.

4 Validation

The construction of PermitTEC v0.1 relies on a natural language processing pipeline to extract case metadata and establish linkages between federal court decisions and the underlying NEPA documents recorded in NEPATEC v2.0. While this automated approach enables processing at a scale that would be impractical through manual effort alone, the policy-sensitive nature of the dataset demands a higher standard of accuracy than automated extraction alone can reliably guarantee. Errors in case metadata, such as incorrect plaintiff and defendant parties, incorrect ruling dates, or inaccurate citations, undermine the analytical integrity of the corpus. More consequentially, incorrect linkages between litigation records and NEPA project entries in NEPATEC v2.0 could lead to inaccurate characterization of which federal actions have faced legal challenge, which agencies bear the highest litigation exposure, and which categories of NEPA documents are most vulnerable to challenge. Because these findings are intended to inform federal permitting policy and agency decision-making, errors of this kind carry risks that extend well beyond academic inaccuracy.

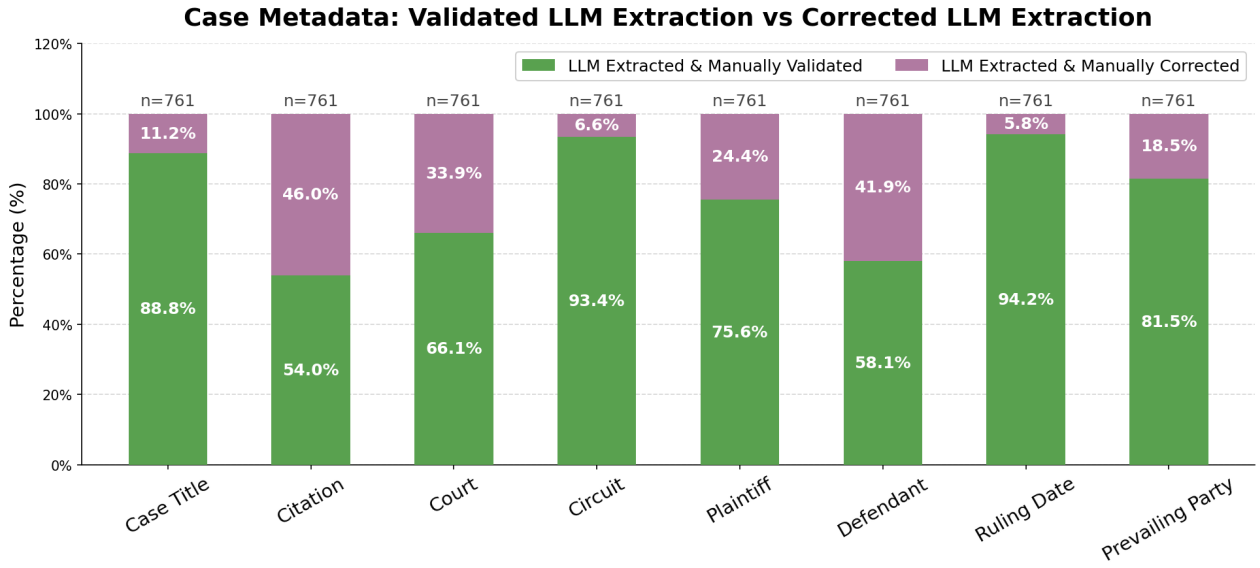


Figure 8. Proportion of LLM-extracted followed by manual validation versus manually corrected LLM-extracted values for each case metadata field across 761 validated litigation records. Blue segments represent fields accepted without modification; orange segments represent fields for which annotators entered a corrected value.

To this end, automated extraction was treated as a first-pass tool, and a structured human-in-the-loop validation process was designed and executed to verify two distinct dimensions of data quality for every record in the dataset: (i) the accuracy of case-level metadata and (ii) the correctness of NEPA project linkages. This section describes each component of this validation effort, the methodology applied, and the outcomes observed.

4.1 Case Metadata Validation

Metadata for each litigation record in PermitTEC v0.1 was initially extracted using an LLM applied directly to the text of the court opinion. Eight structured metadata fields were extracted for each case: Case Title, Citation, Court, Circuit, Plaintiff, Defendant, Ruling Date, and Prevailing Party. These fields form the bibliographic and legal backbone of each record, enabling downstream filtering, aggregation, and linking operations.

Following automated extraction, a structured validation workbook was distributed to a team of trained annotators. Each annotator was assigned a subset of cases and tasked with reviewing the LLM-extracted value for each metadata field against the corresponding source court document. The validation protocol followed a conservative acceptance rule: where the extracted value was correct and complete, the annotator left the corresponding validation cell blank, confirming acceptance of the automated result; while for the entries where the extracted value was incorrect, incomplete, or ambiguous in any respect, the annotator entered a corrected value directly into the workbook, or verifying ambiguous value with a subject matter expert (SME). This design minimized annotation burden while ensuring that every accepted value had been explicitly reviewed by a human expert.

To promote consistency across annotators and reduce subjective variation in judgment, a detailed annotation guide was developed and distributed alongside the workbook prior to the start of the validation exercise. The guide provided field-by-field definitions, formatting requirements, common extraction failure modes, and worked examples drawn from real cases in the dataset. Key conventions enforced through the guide included the following.

Citation. Citations were recorded in standard legal reporter format (e.g., *977 F.3d 853*) derived from the source document filename and confirmed or corrected against the opinion header. Annotators were instructed to record only the volume, reporter, and initial page number, excluding civil action numbers, docket numbers, and parenthetical year references, which the LLM occasionally included in error.

Court and Circuit. Annotators were instructed to record the full institutional name of the issuing court (e.g., *U.S. District Court, District of Nevada*) and to infer the applicable circuit from the court's geographic jurisdiction where not stated explicitly. A common failure mode flagged in the guide was the conflation of the originating district court with the appellate court in cases where the opinion being reviewed was issued on appeal. An additional failure mode involved Supreme Court decisions, which

are not associated with any federal circuit; annotators were instructed to record the circuit for such cases as 'No circuit' rather than inferring a circuit from the geographic origin of the lower court proceedings.

Ruling Date. Dates were required in ISO 8601 format (YYYY-MM-DD). Annotators were directed to locate the ruling date at the conclusion of the opinion, in proximity to the judge's signature or the court's order section, if available. Otherwise, they would use the filing date appearing in the caption at the top of the document.

Prevailing Party. This field was constrained to a controlled vocabulary: *Agency*, *Challenger*, and *Cannot be determined* for cases where there was no clear prevailing party. Cases that did not map cleanly to any standard category were flagged as requiring an SME review.

Plaintiff and Defendant. Full legal names were recorded, with multiple parties separated by semicolons. Annotators were cautioned that in appellate proceedings, the first-listed party may be the appellant rather than the original plaintiff, requiring care to distinguish the original party designations from those arising from the procedural posture of the appeal.

The validated dataset comprises **761 litigation records**. Figure 8 presents the proportion of metadata fields accepted as LLM-extracted versus manually corrected, disaggregated by field type across the full validated corpus.

The results reveal substantial variation in LLM extraction accuracy across field types, reflecting the differing structural predictability of each field within court opinions. Fields with highly standardized presentation — Ruling Date (94.1% accepted), Circuit (93.3% accepted), and Case Title (88.7% accepted) — exhibited the highest rates of automated accuracy, consistent with the relative regularity with which these values appear in court opinion headers and closing sections. Fields requiring greater interpretive judgment showed substantially higher correction rates: Citation required manual correction in 46.1% of cases, Defendant in 42.0%, and Court in 34.0%. The elevated correction rate for Citation reflects the LLM's tendency to include extraneous identifiers such as docket numbers or parenthetical year references rather than recording the standardized reporter citation alone. The high correction rates for Defendant and Court are attributable to the structural complexity of multi-party litigation and the challenge of accurately identifying the precise institutional level of the issuing court, particularly in appellate proceedings where the opinion's caption may reference both the district and appellate courts.

Across all eight metadata fields combined, the LLM-extracted values were accepted without modification in the majority of instances, demonstrating the efficiency of automated extraction as a first-pass tool. At the same time, the correction rates observed across several fields, particularly those requiring legal interpretation, affirm that human review is not merely a precautionary formality but a substantively necessary component of the data production workflow for a corpus intended to support policy-relevant analysis.

4.2 NEPA Project Linkage Validation

Beyond case-level metadata, each litigation record was independently assessed to determine its relationship to the NEPA project documentation recorded in NEPATEC v2.0. This linkage determination is the most consequential step in the PermitTEC v0.1 construction pipeline. As described in Section 1, the challenges associated with linking court decisions to NEPA documents are substantial: court opinions frequently omit or informally reference the underlying NEPA document; some cases challenge the *absence* of a NEPA review rather than the content of an existing one; and automated matching approaches may identify NEPA documents prepared *after* a court decision — often in response to a remand order — as false-positive matches. These challenges make human expert review of linkage determinations essential. Each litigation record was classified into one of three mutually exclusive categories, defined as follows.

Mapped to NEPATEC Project. The litigation challenges a specific NEPA document — an Environmental Impact Statement, Environmental Assessment, Categorical Exclusion, Record of Decision, or Finding of No Significant Impact — that has been identified by the annotator and matched to a corresponding project record in NEPATEC v2.0. To qualify for this classification, the matched NEPA document must predate the court decision, and the linkage must be supported by explicit textual evidence in the opinion.

No NEPA Document Challenged. The litigation does not challenge any NEPA document. This category applies to cases that raise claims under statutes adjacent to NEPA — such as the Clean Water Act, the Endangered Species Act, or the National Historic Preservation Act — without placing a NEPA review directly at issue, as well as to cases in which NEPA is referenced only in passing or in the factual background without forming the basis of any claim. This category also includes cases where the absence of a NEPA process document is being challenged.

NEPA Challenge — Not in NEPATEC. The litigation does challenge a NEPA document, but the underlying project could not be located in NEPATEC v2.0 at the time of validation. This category captures a critical class of cases that are substantively relevant to NEPA litigation analysis but for which the corresponding NEPA record either predates the NEPATEC corpus, was

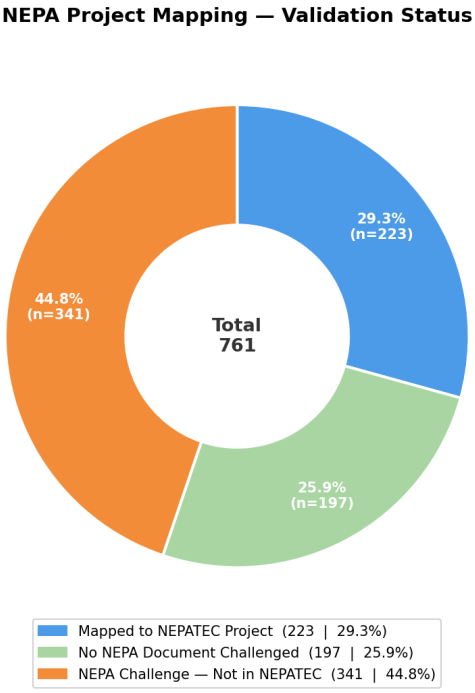


Figure 9. Distribution of NEPA project linkage classifications across 761 validated litigation records.

never digitized, or was otherwise not available in the database. As noted in Section 1, this situation is more common in older cases and in matters involving NEPA document amendments or tiered decisions for which the parent document may exist in NEPATEC while the specific amendment at issue does not.

Annotators made linkage determinations by reviewing the full text of each court opinion, identifying the NEPA document at issue (where applicable), and searching NEPATEC v2.0 for a corresponding project record. The annotation guide provided detailed criteria for distinguishing cases in which a NEPA document is the central object of legal challenge from those in which NEPA is merely referenced, as well as instructions for handling procedurally complex cases involving multiple documents, tiered environmental reviews, and post-remand supplemental analyses. Annotators were further instructed to record the name of the challenged NEPA document, its document type, and its approximate year of issuance where identifiable, and to note cases in which the relevant document could be identified from the opinion text but could not be located in NEPATEC or through publicly available sources.

The validated dataset comprises **761 litigation records** assessed for NEPA project linkage. Figure 9 presents the distribution of linkage classifications across the full validated corpus.

Of the 761 records reviewed, 223 (29.3%) were successfully mapped to a project record in NEPATEC v2.0, representing cases in which the challenged NEPA document was identified and a corresponding project entry was located in the database. A further 197 records (25.9%) were classified as involving no NEPA document challenge — cases that, while part of the environmental litigation landscape and relevant to understanding agency compliance with adjacent statutes, do not directly contest the preparation or adequacy of a federal environmental review. The largest single category, comprising 341 records (44.8%), consists of cases that do challenge a NEPA document but for which a corresponding NEPATEC project entry could not be identified at the time of validation.

This last finding warrants particular attention. That nearly four in ten litigation records in the PermitTEC v0.1 corpus involve NEPA challenges that are not currently captured in NEPATEC v2.0 has two important implications. First, it identifies a concrete opportunity for targeted data expansion: future iterations of NEPATEC could prioritize the acquisition and digitization of NEPA documents associated with cases in this category, many of which involve consequential legal disputes over significant federal actions. Second, it underscores the practical upper bound on the fraction of NEPA litigation that can be linked to NEPATEC project records under current database coverage, a constraint that analysts using the integrated PermitTEC–NEPATEC infrastructure should account for when interpreting corpus-wide statistics.

4.3 Summary of the Validation Task

The manual validation effort described in this section encompassed 761 litigation records and applied structured human expert review to two complementary dimensions of data quality: the accuracy of case-level metadata and the correctness of NEPA project linkages. The combination of LLM-assisted extraction and structured human review proved effective across both validation tasks. Automated extraction provided a reliable baseline for the majority of metadata fields while human annotators identified and corrected systematic failure modes, particularly in fields requiring legal interpretation, that would otherwise have propagated into the final corpus undetected. For NEPA project linkages, human review was indispensable: the nuanced distinctions between cases that challenge NEPA documents, cases that challenge the absence of such documents, and cases that reference NEPA only peripherally cannot be reliably resolved through automated classification alone, and the consequences of misclassification for downstream policy analysis are significant.

The resulting dataset reflects a standard of accuracy and provenance traceability appropriate for use in policy analysis, regulatory research, and reporting intended for federal decision-makers. Every metadata field in PermitTEC v0.1 is explicitly tagged to indicate whether its value was accepted from automated extraction or entered by a human annotator, providing full transparency about data provenance and enabling downstream users to calibrate confidence appropriately depending on the sensitivity of their application.

5 Usage

The PermitTEC v0.1 litigation corpus, combined with its linkages to NEPA project records in NEPATEC v2.0, supports a range of analytical and applied use cases spanning environmental law, permitting practice, and computational legal research. While some applications are enabled directly by the current release, others will become fully realizable as the corpus expands and paginated text content is incorporated in future releases.

Litigation Pattern Analysis. The structured metadata enables systematic analysis of NEPA litigation patterns across temporal, geographic, agency, and statutory dimensions. Researchers can examine trends in litigation volume, identify jurisdictional variation in disposition outcomes across federal circuits, and determine which agencies face the highest rates of challenge and under which statutes. The three-category litigation classification further supports disaggregated analysis of how challenges to specific NEPA documents, the absence of required reviews, and adjacent-statute claims are distributed across the corpus.

Project-Level Risk Assessment. Attributes such as project type, sector, lead agency, geographic location, and NEPA document type — drawn from the linked NEPATEC v2.0 project records — can be correlated with litigation incidence and disposition outcomes to identify combinations of characteristics associated with elevated legal risk. For proposed or early-stage federal actions, practitioners can compare a project's attributes against these historical patterns to assess the likelihood and nature of potential challenges, enabling precedent-informed refinements to draft NEPA documents, alternatives analyses, and public engagement strategies before the review process is completed.

Cross-Corpus Defensibility Analysis. Building on these risk profiles, the corpus supports more granular investigation of what makes a NEPA document legally defensible. By examining cases where courts found compliance adequate (e.g., summary judgment for federal defendant) alongside cases where deficiencies were identified (e.g., remand with vacatur, injunction issued), researchers can study whether NEPA document type, analytical depth, or agency practice correlates with litigation resilience — and systematically quantify the substantive and procedural factors that distinguish successful from unsuccessful environmental reviews across agencies, project types, andjurisdictions.

AI-Assisted Legal Research. The corpus provides a structured foundation for building retrieval systems that support targeted precedent search based on project characteristics, agency, jurisdiction, or legal grounds rather than keyword search alone. As text content is added in future releases, more advanced capabilities become feasible: automated summarization of judicial reasoning with source attribution, extraction of holdings, and identification of specific NEPA analyses found deficient by courts.

NLP Benchmarking. The corpus serves as a resource for training and evaluating models on legal NLP tasks,including litigation type classification, named entity recognition in judicial text, cross-document entity resolution between informal project references and canonical records, and disposition prediction. The 329-case benchmark with ground truth across nine metadata fields provides a standardized evaluation set for legal information extraction in the environmental domain.

Policy Analysis. Temporal and jurisdictional metadata support empirical analysis of how regulatory changes — such as updates to CEQ regulations or agency-specific NEPA procedures — affect litigation volume, challenge types, and disposition distributions over time, informing evidence-based approaches to permitting reform.

Corpus Maintenance. The automated ETL pipeline described in this work is designed to support ongoing corpus expansion. The ingestion, extraction, and classification infrastructure can be configured to monitor federal court dockets for new NEPA-related decisions and integrate them with minimal manual intervention, maintaining the corpus as a living resource rather than a static snapshot.

6 Limitations

We note several known limitations of our PermitTEC v0.1 dataset.

First, PermitTEC inherits a coverage gap from project-level linkage to NEPATEC 2.0. As reflected in the linkage statistics above, 41.9% of litigation records involve NEPA challenges for which the underlying project is not currently represented in NEPATEC 2.0. This occurs most often for older cases and for challenges to NEPA document amendments or tiered decisions. Users should account for this constraint when computing corpus-wide statistics on matched cases or when interpreting analyses that depend on complete project-level linkage.

Second, some structured fields were generated through LLM-assisted extraction. Although the dataset includes human validation, the source tag allows users to identify fields that were not manually reviewed. In particular, fields tagged `llm_extracted_no_manual_review` should be treated with greater caution in precision-sensitive applications.

Third, party designation may require care in appellate matters. In appellate proceedings, the first-listed party in the case caption may be the appellant rather than the original plaintiff. To maintain consistency across the dataset, party fields reflect the original plaintiff and defendant designations rather than the appellate ordering.

Fourth, Supreme Court cases carry *No circuit* in the `circuit` field, since they are not associated with any federal circuit.

Finally, some cases involve challenges to multiple NEPA documents. In these instances, the `linked_to` block reflects the primary document identified during validation, and secondary documents may not be fully captured in PermitTEC v0.1.

Acknowledgments

This work was supported by the Office of Critical Minerals and Energy Innovation (CMEI), U.S. Department of Energy, and was conducted at Pacific Northwest National Laboratory, which is operated by Battelle Memorial Institute for the U.S. Department of Energy under Contract DE-AC05–76RL01830.

We would like to thank Jordan Eccles from the White House Council on Environmental Quality (CEQ), Pranava Raparla and Jack Titus from the U.S. Department of Energy Office of Policy, and all others who contributed to the discussion in developing PermitTEC v0.1 metadata. We would also like to thank the Breakthrough Institute (BTI) for providing the valuable case documents to aid with this dataset construction.

This technical report has been cleared by PNNL for public release as PNNL-39193.

References

1. Lightbody, L. Winter v. Natural Resources Defense Council, Inc. *Harv. Environ. Law Rev.* **33**, 593 (2009).

2. Munikoti, S. *et al.* NEPATEC v2. 0: Standardized Metadata and Text Corpus of National Environmental Policy Act Documents. Tech. Rep., Pacific Northwest National Laboratory (PNNL), Richland, WA (United States) (2025). DOI: 10.2172/2584716.

3. Chiappa, N., Nordhaus, T., Trembath, A., McCarthy, E. & Hernandez, J. Understanding NEPA Litigation: A Systematic Review of Recent NEPA-Related Appellate Court Cases. Tech. Rep., The Breakthrough Institute (2024). Accessed: 2026-03-30.

4. Thomson Reuters. Westlaw (2026). Accessed: 2026-03-30.

5. Free Law Project. Courtlistener (2026). Accessed: 2026-03-30.

6. Khattab, O. *et al.* Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714* (2023).

7. Google. Gemini API documentation: Interactions (2026). Accessed: 2026-03-30.

A Automated ETL Pipeline for Litigation Document Ingestion and Processing

To operationalize the individual components of the PermitTEC v0.1 workflow as a reproducible and scalable system, we developed an automated extract-transform-load (ETL) pipeline that orchestrates the full document processing lifecycle — from raw document ingestion through AI-driven metadata extraction to integration into a searchable data lakehouse following the PermitTEC schema (Figure 5).

The pipeline employs a hybrid multi-cloud architecture in which data authority and document lifecycle management reside on AWS, while compute-intensive AI/ML metadata extraction is executed on GCP. This separation decouples document storage, orchestration, and provenance tracking from model-driven processing, allowing each layer to scale independently. On AWS, incoming documents are ingested into Amazon S3, where event-driven notifications trigger an SQS-based orchestration workflow. Text extraction is performed using a PyMuPDF-based batch processor with quality validation and confidence scoring. Extracted text and relational metadata are persisted into PostgreSQL — extending the NEPATEC schema — while OpenSearch indexes both text and metadata to support full-text search and analytics across the corpus. DynamoDB maintains end-to-end provenance records for every processing step. On GCP, extracted text staged by a cross-cloud transfer function is received into Cloud Storage and routed to Vertex AI-based processors for metadata extraction using large language models, with dynamic routing based on document complexity and quality characteristics.

Although the current release processes NEPA litigation documents, the pipeline architecture is designed to be document-type agnostic. The PermitTEC schema and processing modules are structured to accommodate additional federal and state environmental permitting document types in future releases, including Clean Water Act permitting documents, Endangered Species Act consultation records, National Historic Preservation Act compliance documents, state-specific environmental permitting documents, and geothermal or critical minerals project permitting records. This extensibility is achieved through a modular design in which document-type-specific extraction prompts, metadata schemas, and quality validation rules can be configured independently of the core ingestion and orchestration infrastructure.

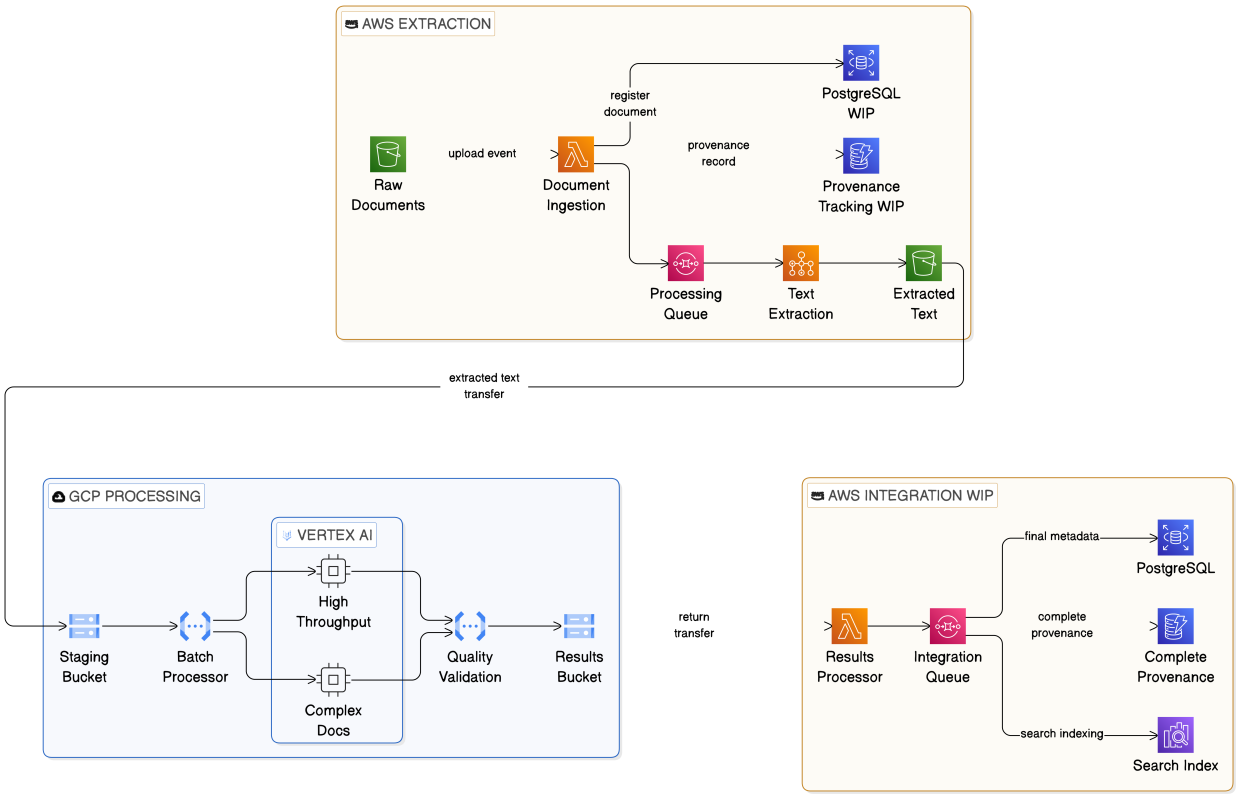


Figure 10. High-level system architecture showing the hybrid multi-cloud implementation. AWS manages document ingestion, text extraction, relational data, search indexing, and provenance tracking; GCP provides AI/ML-driven metadata extraction and quality validation.

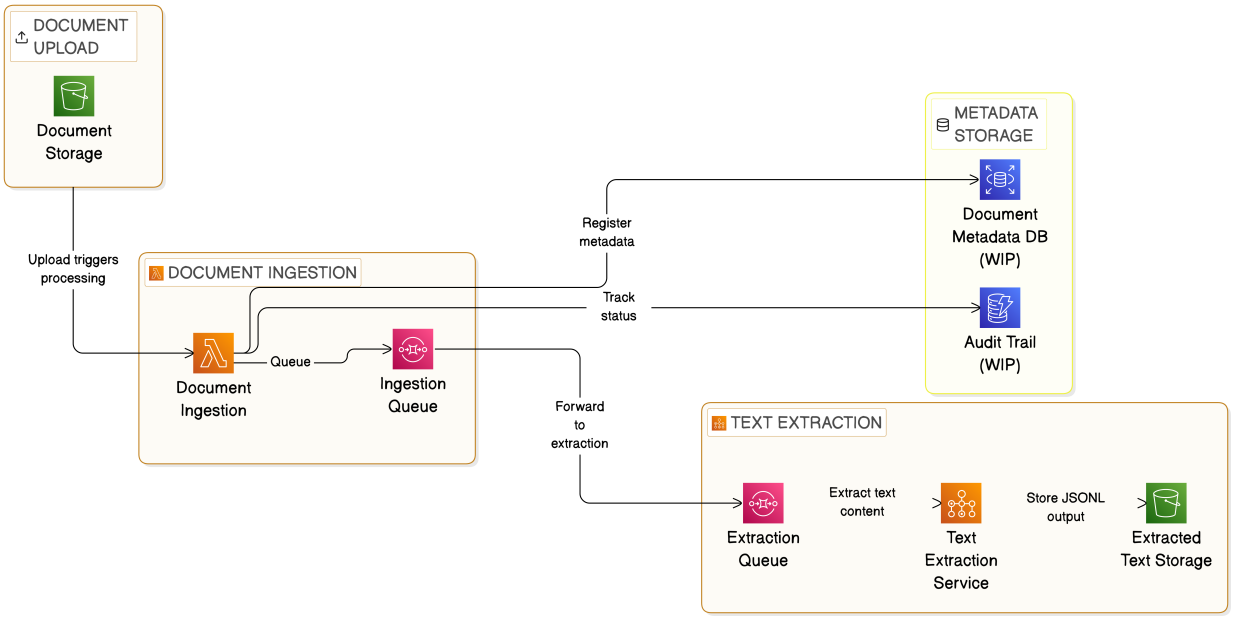


Figure 11. AWS-side processing pipeline.

A.1 System Architecture

The ETL pipeline is implemented using a hybrid multi-cloud architecture that separates data authority and document lifecycle management from compute-intensive AI/ML processing. Figure 10 provides a high-level overview of the cross-cloud implementation. AWS serves as the authoritative hub for document ingestion, text extraction, relational data management, search indexing, and provenance tracking, while GCP functions as the dedicated AI/ML processing engine for metadata extraction and advanced document intelligence. This separation decouples data management from model-driven workloads, enabling each layer to scale independently and allowing best-in-class services to be selected across cloud providers.

The architecture is organized around four design principles: (1) *permitting standardization* through unified data models across diverse federal and state regulatory frameworks; (2) *AI-native processing* through large language model metadata extraction and relationship discovery; (3) *cross-jurisdictional intelligence* enabling interoperability across agencies and permitting regimes; and (4) a *scalable data lakehouse* that unifies raw documents, extracted text, structured metadata, and provenance into a single searchable system.

Primary Data Authority and Processing (AWS). The AWS environment is the primary entry point for documents entering the pipeline (Figure 11). Incoming documents are stored in Amazon S3, where event notifications trigger an SQS-driven orchestration workflow. Text extraction is performed using a PyMuPDF-based batch processor that parses document content and computes quality and confidence metrics characterizing the reliability of extracted text for downstream processing. Extracted text and associated metadata are written to processed S3 buckets, while relational data — including document relationships extending the NEPATEC schema — is persisted into PostgreSQL. OpenSearch indexes both text and metadata to support full-text search and analytics across the corpus, and DynamoDB maintains provenance records for each processing step. Together, these components provide the system of record for document storage, text extraction and preprocessing, search indexing, and infrastructure orchestration.

AI/ML Metadata Extraction (GCP). Extracted text staged by the AWS cross-cloud transfer function is received into GCP Cloud Storage input buckets, organized by tenant and execution ID (Figure 12). A Cloud Functions-based orchestrator evaluates each document's structural complexity, text quality, and agency-specific requirements to dynamically route batches to the appropriate model endpoint. Standard documents are routed to Gemini 2.5 Flash for high-throughput, cost-efficient extraction, while complex legal, multi-jurisdictional, or low-quality documents are escalated to Gemini 2.5 Pro for higher extraction accuracy. Vertex AI Batch API manages job execution across configurable concurrent batches, with results validated through a multi-level confidence scoring framework that evaluates text extraction quality, field-level extraction confidence, schema compliance, cross-validation consistency, and model prediction certainty. Documents that fail to meet agency-specific confidence thresholds are automatically reprocessed with escalated model selection. Beyond single-document metadata extraction, the GCP layer

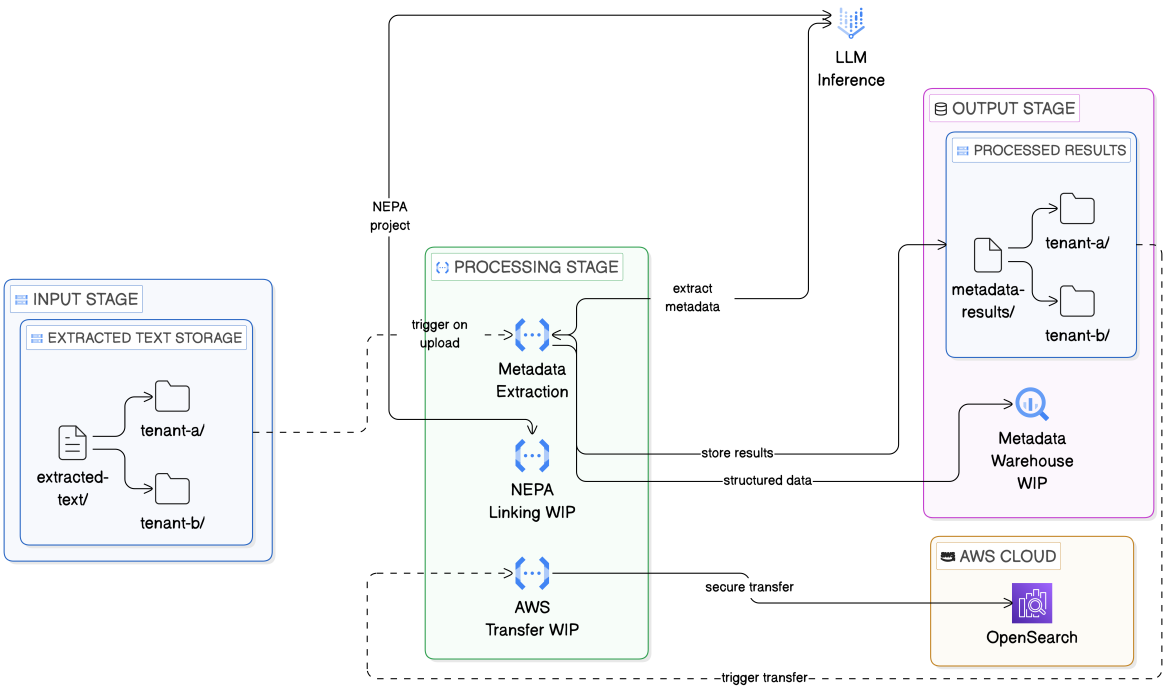


Figure 12. GCP-side AI/ML processing pipeline.

also performs multi-document relationship analysis, discovering temporal, hierarchical, cross-reference, and geographic relationships across document clusters. Final validated results are transformed to PermitTEC schema compliance and delivered back to AWS for integration into the data lakehouse. In this design, the GCP environment performs no persistent data storage; it functions as a stateless intelligence layer for scalable, quality-assured AI/ML inference.

A.2 Data Provenance

Reproducibility and auditability are essential properties of a document processing pipeline operating across multiple cloud environments, document types, and regulatory jurisdictions. As documents move through ingestion, text extraction, AI-driven metadata enrichment, and final integration, each processing decision, intermediate output, and quality assessment must be traceable to support debugging, compliance verification, and incremental reprocessing. To meet these requirements, the pipeline implements a dedicated provenance layer that provides end-to-end lineage tracking across both AWS and GCP. Figure 13 illustrates how provenance is tracked through successive pipeline stages.

The provenance model is organized as a three-tier hierarchy capturing lineage at the execution, chunk, and document levels. At the top level, *execution records* capture job-level metadata including initialization parameters, processing configuration, status transitions, chunk counts, and cross-cloud context. At the intermediate level, *chunk records* track batch-level processing details including storage locations, processing stage timelines, and per-chunk quality metrics. At the most granular level, *document records* maintain the full processing history of each individual PDF, identified by a content-based hash that enables duplicate detection across executions. Together, these three tiers support complete reconstruction of a document's path through the pipeline, including timing, quality metrics, and generated outputs. The corresponding Pydantic-based schemas are provided in Appendix D.

The provenance layer is implemented through an event-driven Lambda function that responds to six lifecycle events from the pipeline's Step Functions orchestrator: execution initialization, chunk creation, stage completion, failure handling, execution finalization, and provenance query. All provenance records are persisted in a DynamoDB-backed store with Pydantic-enforced validation on every write, ensuring schema consistency across the pipeline. This provenance infrastructure is tightly coupled with the queue-based architecture: each SQS message carries tracking fields such as execution ID, document hash, tenant ID, and current stage, enabling fine-grained tracing and automated retry logic. Cross-cloud lineage between AWS and GCP is maintained through dedicated context fields and transfer audit events, while tenant isolation is enforced via S3 prefixes, queue attributes, and DynamoDB partitioning.

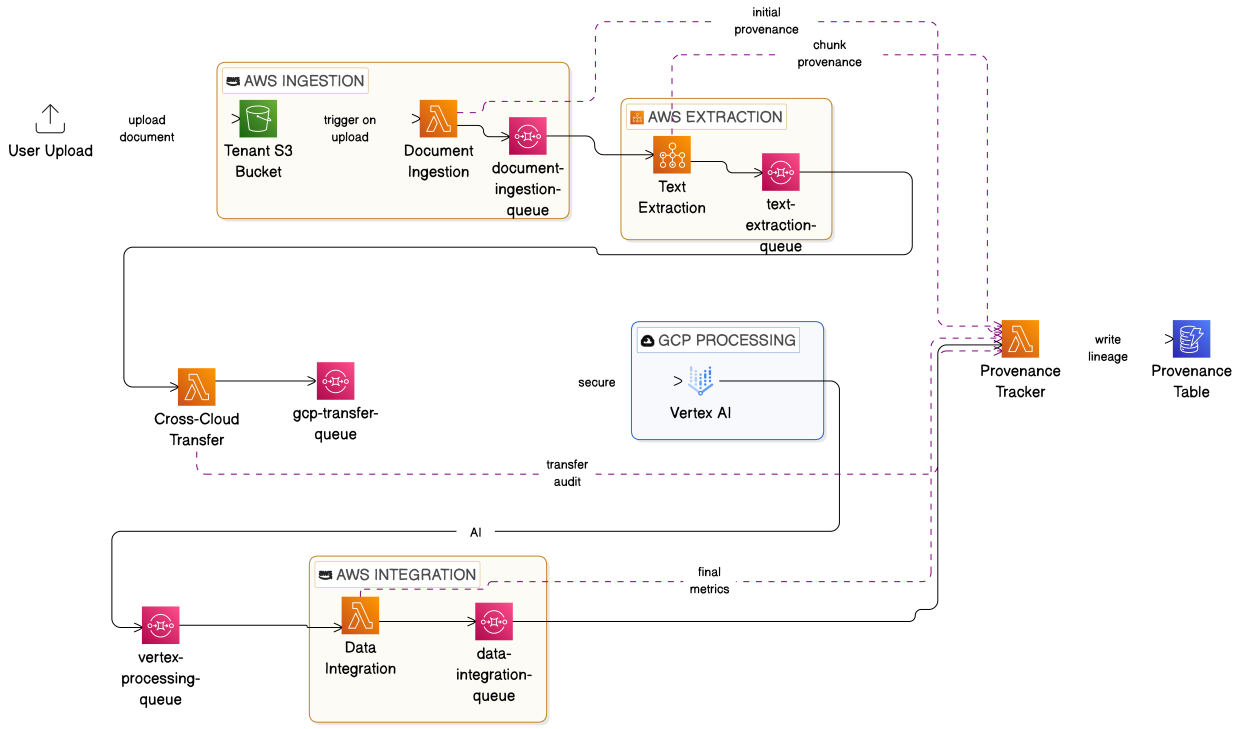


Figure 13. Cross-cloud provenance architecture tracking document status through successive pipeline stages.

B Example entry of PermitTEC v0.1


```
{
  "case_uuid": "case-2-f-4th-953",
  "case_metadata": {
    "case_title": {
      "value": "ENVIRONMENTAL DEFENSE FUND v. FEDERAL ENERGY REGULATORY COMMISSION, Spire
      ↪  Missouri Inc., and Spire STL Pipeline LLC",
      "source": "llm_extracted_and_manually_validated"
    },
    "citation": {
      "value": "2 F. 4th 953",
      "source": "llm_extracted_and_manually_validated"
    },
    "court": {
      "value": "United States Court of Appeals, District of Columbia Circuit",
      "source": "llm_extracted_and_manually_validated"
    },
    "circuit": {
      "value": "District of Columbia",
      "source": "llm_extracted_and_manually_validated"
    },
    "plaintiff": {
      "value": "Environmental Defense Fund",
      "source": "llm_extracted_and_manually_validated"
    },
    "defendant": {
      "value": "FEDERAL ENERGY REGULATORY COMMISSION",
      "source": "llm_extracted_and_manually_validated"
    },
    "ruling_date": {
      "value": "2021-06-22",
      "source": "llm_extracted_and_manually_validated"
    },
    "prevailing_party": {
      "value": "Challenger",
      "source": "llm_extracted_and_manually_validated"
    }
  },
  "linked_to": {
    "in_nepatec": true,
    "llm_extracted_keywords": [
      "Spire STL Pipeline",
      "FERC",
      "Natural Gas Pipeline",
      "St. Louis",
      "Certificate of Public Convenience"
    ],
    "llm_extracted_title": "Spire STL Pipeline",
    "contested_project_name": "Spire STL Pipeline Project",
    "nepatec_project_uuid": "ddee6a29a3a06a44f495850065fcd664"
  }
}
```


Listing 1. Example JSON representation of a PermitTEC v0.1 record.

C Instruction Prompts Used

C.1 Prompts used in Metadata Extraction

Instruction Prompt For Extracting Titles/Abstract from Litigation Documents

You are given a litigation document (PDF) that contains a formal court case opinion or filing. Text is obtained from a default PDF reader extraction. Your task is to extract the **contested project name excerpt** from the litigation document. This contested project name excerpt corresponds to a section of the text that mentions which action the litigation was brought against.
Instructions:

- The contested project name typically is mentioned in the Background, Facts, or Introduction sections of the litigation document where the initial reason for the case being brought to court is mentioned.
- The contested project name may be mentioned with the word "Project" as an identifier, or a general description of the action taken by an agency is mentioned (e.g., "The feasability study to reduce the flood risk in the Watson Metropolitan Area")
- The contested project name may be mentioned with reference to a specific analysis or assessment required by the National Environmental Policy Act (NEPA) (e.g., EIS, FEIS, EA, BO, ROD, FONSI)
- The government entity/agency may be referenced in regard to whom was conducting or initiating the action.
- The contested project name is referenced as a past action and having occurred in the past.
- The location of the project name will typically be mentioned when referencing the specific action.

The extracted contested project name excerpt should:

- Be a couple of sentences (maximum of 4), directly extracted from the text, describing the issue bringing forth litigation.
- Include the specific action the litigation was brought against (e.g., the construction of [building], the assessment of [action])

Examples of valid contested project name excerpts:

- Plaintiffs allege that the FAA violated the NEPA through its role in approving projects at PHL. including the Philadelphia International Airport Runway 17-35 Improvements Project ("Runway 17-35 Project").
- On April 20, 2010, the Deepwater Horizon, a deep-water exploratory oil rig, exploded, caught fire, and sank in the Gulf of Mexico, resulting in the largest oil spill in the United States in modern history. Less than two years later, the Bureau of Ocean Energy Management (“BOEM”), approved two lease sales in the area where the Deepwater Horizon spill occurred. The plaintiffs bring this action challenging BOEM’s approval of those lease sales under the National Environmental Policy Act, the Administrative Procedure Act, and the Endangered Species Act.
- Alabama Power operates seven hydroelectric generator and storage developments along waterways located primarily in Alabama. The Company’s developments on the Coosa River (“the Coosa Project”) are at the center of this dispute.

Instruction Prompt For Extracting Geographic State from Litigation Documents

You are given a litigation document (PDF) that contains a formal court case opinion or filing. Text is obtained from a default PDF reader extraction. Your task is to extract the **state** which identifies the location where the case was initially brought to court. Instructions:

- The state typically appears on the first page, or first few pages, of the documents.
- For district level cases, the state often appears in the name of the district court. The state may also be found in the introduction section of the document or within the text of the document when the background of the case and the action causing the case to be brought to court is discussed. If the state included in the district court name is different from the state(s) mentioned in the discussion on the cases context, include all references states separated by semi-colons (e.g., NC; NY; MA)
- For appellate level cases, the state may be found in the introduction section of the document or within the text of the document when the background of the case and the action causing the case to be brought to court is discussed. If the state (or states) is not explicitly mentioned in the cases context, list all of the states/territories included in the appeals circuit, separated by semi-colons. (e.g. for 11th Circuit: AL; FL; GA)
- For appellate level cases, below is a mapping of states/territories to circuits:
  - 1st Circuit: Maine, Massachusetts, New Hampshire, Puerto Rico, Rhode Island
  - 2nd Circuit: Connecticut, New York, Vermont
  - 3rd Circuit: Delaware, New Jersey, Pennsylvania, Virgin Isla
  - 4th Circuit: Maryland, North Carolina, South Carolina, Virginia, West Virginia
  - 5th Circuit: Louisiana, Mississippi, Texas
  - 6th Circuit: Kentucky, Michigan, Ohio, Tennessee
  - 7th Circuit: Illinois, Indiana, Wisconsin
  - 8th Circuit: Arkansas, Iowa, Minnesota, Missouri, Nebraska, North Dakota, South Dakota
  - 9th Circuit: Alaska, Arizona, California, Guam, Hawaii, Idaho, Montana, Nevada, Northern Mariana Islands, Oregon, Washington
  - 10th Circuit: Colorado, Kansas, New Mexico, Oklahoma, Utah, Wyoming
  - 11th Circuit: Alabama, Florida, Georgia

The extracted state should:

- List at least one state/territory.
- When more than one state/territory is listed, separate by semi-colons;
- Be the abbreviation of the state (e.g. New York = NY, Washington = WA)
- If the location is a territory of the United States, and not one of the 50 states, write the full name out (e.g., Puerto Rico, Northern Mariana Islands)

Examples of valid states:

- NY
- NC; VA; WV; SC
- D.C.
- Northern Mariana Islands
- Alaska, Arizona, California, Guam, Hawaii, Idaho, Montana, Nevada, Northern Mariana Islands, Oregon, Washington
- Puerto Rico

C.2 Prompts used in Litigation to NEPA Mapping

Prompt used in Method 1 for Extracting Project Title and Relevant Query Keywords

You are an expert environmental policy analyst specializing in NEPA (National Environmental Policy Act) documents and litigation.
TASK: Analyze this litigation document (PDF) to identify the underlying NEPA project that was challenged in this lawsuit. The litigation was filed AFTER the NEPA documents were released, so you need to identify the original federal project/action that this case is about.
IMPORTANT CONTEXT: - Litigation documents reference NEPA documents/projects that were already created - The NEPA project is the underlying federal action being challenged (pipeline, mining, land use, etc.) - You should use web search to find additional context about this NEPA project if needed
INFORMATION TO EXTRACT:
1. **Project Title**: The official or commonly used name of the underlying NEPA project/action - Look for project names, action descriptions, or federal proposals mentioned - This should be the NEPA project name, NOT the litigation case name
2. **Search Keywords** (exactly 5): Keywords that would help find this specific NEPA project in a document database - Include: project name components, geographic locations, agency names, action types, unique identifiers - Prioritize: specificity over generality - Format: distinct, searchable terms
EXAMPLES OF GOOD OUTPUT:
Project Title: "Haiwee Geothermal Leasing Area Environmental Assessment" Keywords: ["Haiwee Geothermal", "BLM", "Inyo County California", "geothermal leasing", "environmental assessment"]
Project Title: "Dakota Access Pipeline Environmental Impact Statement" Keywords: ["Dakota Access Pipeline", "DAPL", "Army Corps Engineers", "North Dakota", "crude oil pipeline"]
OUTPUT FORMAT: Provide your response as a JSON object with exactly this structure:
{
“project_title":“Official or common name of the NEPA project",
“keywords": “keyword1", “keyword2", “keyword3", “keyword4", “keyword5"],
“confidence": “high|medium|low",
“reasoning": “Brief explanation of how you identified this project"
}

INSTRUCTIONS: 1. Read and understand the litigation PDF thoroughly 2. Use Google Search if needed to find more context about the NEPA project 3. Identify the underlying NEPA project that is being challenged 4. Extract the project title and 5 most relevant search keywords 5. Assess your confidence level based on clarity of information 6. Provide reasoning for your identification Analyze the PDF now and return the JSON response:

D Data Provenance Record Schemas

This appendix presents the Pydantic-based data models for the three-tier provenance hierarchy described in Section A.2. Each model defines the schema for provenance records persisted in DynamoDB, enforcing type validation and structural consistency across all pipeline stages.

D.1 Execution Record Schema

The execution record captures job-level metadata for each pipeline run, including processing configuration, status transitions, chunk and document counts, and cross-cloud context. Each record is uniquely identified by `execution_id` with multi-tenant isolation.

```
class ExecutionRecord(BaseModel):
    execution_id: str
    sort_key: Literal["METADATA"]
    record_type: Literal[RecordType.EXECUTION]
    status: ProcessingStatus
    created_at: datetime
    updated_at: datetime
    completed_at: datetime | None
    agency_context: AgencyContext
    processing_config: ProcessingConfig
    execution_summary: dict
    total_chunks: int
    total_pdfs: int
    completed_chunks: int
    failed_chunks: int
    cross_cloud_context: CrossCloudContext | None
    step_function_arn: str
    execution_arn: str
    ttl_timestamp: int
```

D.2 Chunk Record Schema

The chunk record tracks batch-level processing details for each group of documents processed together, including S3 storage locations, processing stage timelines, and per-chunk quality and timing metrics.

```
class ChunkRecord(BaseModel):
    execution_id: str
    sort_key: str  # CHUNK#{chunk_id}
    record_type: Literal[RecordType.CHUNK]
    chunk_id: str
    chunk_index: int
    payload_s3_key: str
    status: ProcessingStatus
    created_at: datetime
    updated_at: datetime
    started_processing_at: datetime | None
    completed_at: datetime | None
    pdf_count: int
    pdf_files: list[str]
    pdf_hashes: list[str]
    processing_path: str
    current_stage: str
    stage_timeline: list[StageTimeline]
    total_processing_time_seconds: float | None
    avg_pdf_processing_time: float | None
    ttl_timestamp: int
```

D.3 Document Record Schema

The document record maintains the complete processing history of each individual PDF, identified by a content-based hash (`pdf_hash`) that supports duplicate detection across executions. The record tracks processing stages, output artifacts, and references to generated OpenSearch and vector index entries.

```
class PDFRecord(BaseModel):
    execution_id: str
    sort_key: str  # PDF#{pdf_hash}
    record_type: Literal[RecordType.PDF]
    pdf_hash: str
    filename: str
    original_s3_path: str
    chunk_id: str
    status: ProcessingStatus
    created_at: datetime
    updated_at: datetime
    is_duplicate: bool
    duplicate_of_hash: str | None
    first_seen_execution: str | None
    duplicate_count: int
    processing_history: list[StageTimeline]
    opensearch_records: list[str]
    vector_records: list[str]
    output_artifacts: dict[str, str]
```

E Workflows for Litigation to NEPA Mapping

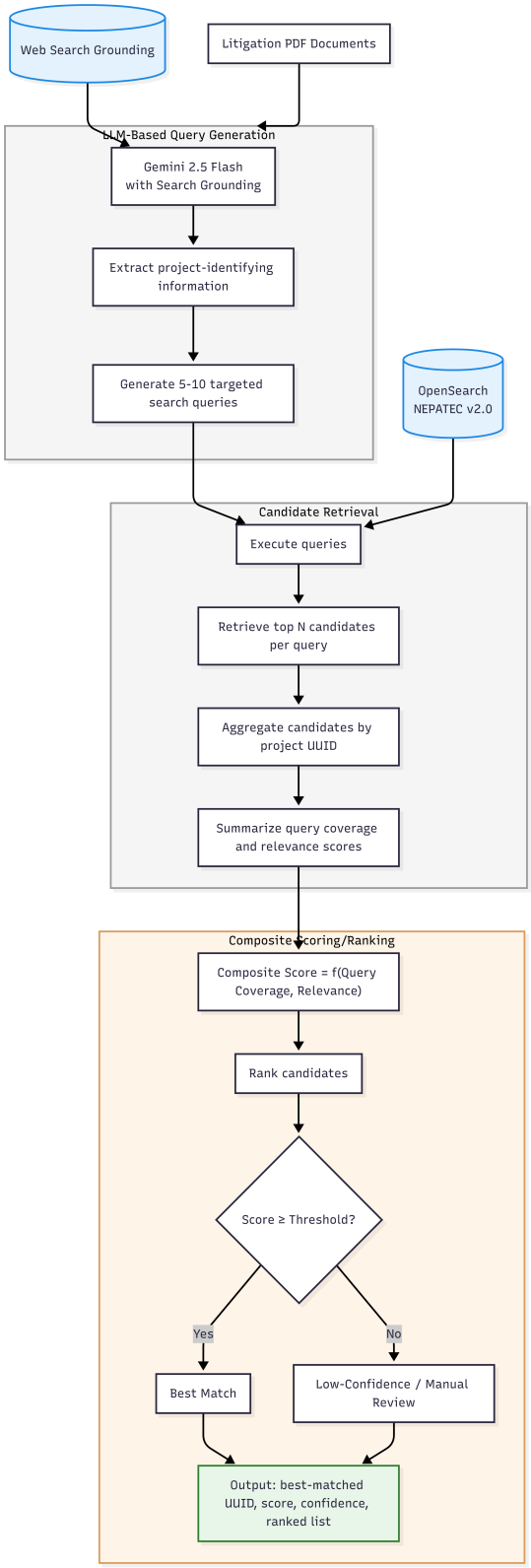


Figure 14. Method 1 schematic illustrating search-grounded query generation and retrieval workflow.

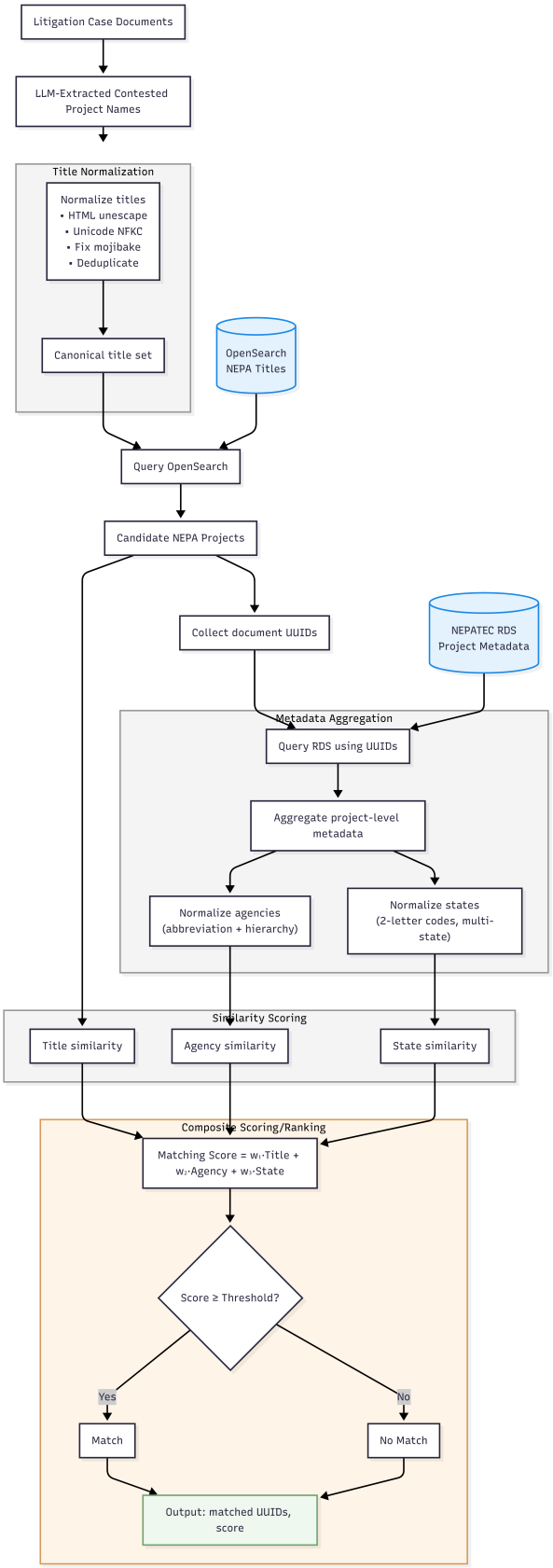


Figure 15. Method 2 schematic illustrating the title fuzzy matching workflow with composite scoring based on title, agency, and state.

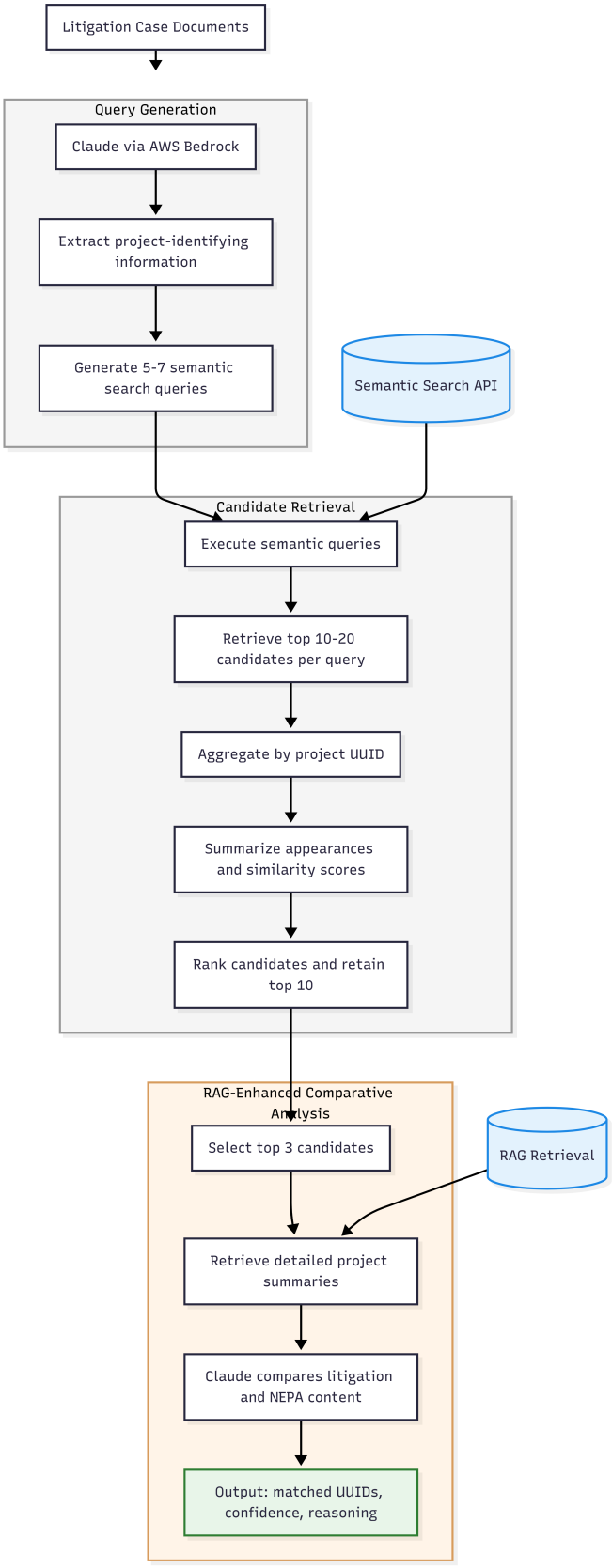


Figure 16. Method 3 schematic illustrating semantic retrieval and RAG-enhanced comparative analysis workflow.