Technical Report & Recommendations

Phase 1: NEPA Decarbonization Technology Analysis

July 31, 2026

Table of contents

Download PDF version

Executive Summary

This report is a technical companion to the six Phase 1 deliverable reports. It explains how the analysis was built — the dataset, the extraction pipelines, the validation performed — and what someone would need to know to replicate or extend the work. Findings live in the deliverable reports; this document is about the machinery behind them.

Phase 1 was the first analysis layer built on NEPATEC 2.0, a corpus of more than 120,000 NEPA documents from roughly 60,000 federal reviews. From that corpus, Phase 1 classified a 20,725-project decarbonization universe and delivered six analyses: the technology, agency, and geographic landscape of decarbonization reviews; the prevalence and speed of programmatic and tiered reviews; review-type shares, timelines, and generation capacity; multi-state and multi-department collaboration structure; document length and compliance with the Fiscal Responsibility Act's page limits; and technology-specific deep dives on transmission, geothermal, and pipelines.

None of the quantities these analyses needed — review dates, capacity figures, page counts net of appendices, collaboration roles — exist as structured data in the source. Everything was extracted from document text with a common pattern: metadata first, then capped page scans of main documents, then regular expressions with surrounding context, then trained text classifiers where the volume required them, and a large language model only for the small residue of ambiguous cases. Estimated spend on language-model adjudication across all of Phase 1 was on the order of five dollars.

Phase 1's validation was manual and targeted — spot-check samples, case-level review memos, client validation packets — rather than the formal held-out gold sets introduced in Phase 2; the sections below note which numbers rest on which kind of check. The report closes with recommendations about the dataset itself: the structured metadata, record linkage, and publication practices that would have made most of this extraction unnecessary, and that would make every future analysis of federal permitting less expensive and more reliable.

The Dataset

Phase 1 is built on NEPATEC 2.0, assembled by Pacific Northwest National Laboratory's PermitAI project and published on Hugging Face. The corpus contains more than 120,000 NEPA documents from roughly 60,000 federal reviews conducted by more than 60 agencies — millions of pages of text. It is organized as three separate collections by review type: Categorical Exclusions (CE), Environmental Assessments (EA), and Environmental Impact Statements (EIS), each stored as nested machine-readable records (project → process → documents → pages) with project-level metadata.

Phase 1's extraction pipeline flattens this into a single analysis table, `projects_combined.parquet` — one row per project, 61,881 rows and 97 columns in the frozen version, carrying the original metadata plus every field Phase 1 extracted. The full inventory breaks down as follows:

Energy classification	CE	EA	EIS	Total
Decarbonization	19,399	573	753	20,725
Fossil fuel	9,191	969	623	10,783
Other	26,078	1,541	2,754	30,373
Total	**54,668**	**3,083**	**4,130**	**61,881**

The decarbonization universe — the subject of all six deliverables — was defined by NEPATEC's own project-type tags: a project counts as decarbonization when it carries at least one decarbonization tag (renewable energy production, electricity transmission, nuclear technology, carbon capture and sequestration, utilities, and related categories) and no fossil-fuel tag (coal, land-based or offshore oil and gas, pipelines, rural energy). Because tags are broad, three documented refinements then removed projects the tags capture but the analysis should not: utility-tagged projects that are actually broadband, waste, or land development (1,623 removed); military and defense nuclear actions (481); and nuclear-waste cleanup work at DOE environmental-management sites, filtered by office and site lists with a client-reviewed keep list of 34 genuinely energy-relevant projects. The full logic, term lists, and counts are documented in the repository and summarized in the project overview.

One structural fact about NEPATEC shapes every Phase 1 result: EIS records are drawn from EPA's governmentwide EIS database and are broadly comprehensive since 2012, but EA and CE records are substantially complete only for the Department of Energy and the Bureau of Land Management. DOE alone accounts for 81 percent of decarbonization projects, 96 percent of them categorical exclusions. Statistics about EAs and CEs therefore describe the DOE/BLM-dominated corpus, not the full federal population.

Pipeline Architecture

Phase 1 splits cleanly by language: Python for extraction (about 17,700 lines across thirteen extraction scripts), R and the tidyverse for analysis and figures, and Quarto for the published reports. Bulk page text is read through DuckDB rather than loaded into memory — the EIS page store alone is roughly 5.5 GB — using predicate-pushdown queries that pull only the pages a given extraction needs, capped per project (typically the first 50–80 pages of main documents).

Every extractor follows the same four-step design pattern. First, check project metadata — many quantities appear in titles or descriptions and need no document reading at all. Second, scan a capped number of pages from documents flagged as main documents, using regular expressions that capture the matched value plus surrounding context. Third, apply deterministic rules to score and select among candidates — pattern strength, document-type priority, position on the page, and false-positive filters tuned to known traps. Fourth, and only when multiple plausible candidates survive the rules, send the candidates with their context to a language model for adjudication. Every stage writes an audit trail: which method produced the value, what the alternatives were, and — when a model was involved — the model name, its reasoning, and its confidence.

Three kinds of models appear in the pipeline. Fine-tuned BERT-family classifiers handle high-volume text classification: date-context classification for timelines and a science-domain classifier (SciBERT) for geothermal project phases, trained with class-weighted loss and self-training on the pipeline's own high-confidence predictions. Claude Haiku handles low-volume adjudication — choosing among capacity candidates, selecting the right dates for complex EA/EIS timelines, resolving transmission length conflicts. The scale is small: an internal cost estimate puts total Phase 1 language-model spend at roughly $5 across a few thousand calls (the one component with a measured actual came in near twice its estimate, so the true total is likely somewhat higher, but of the same order). The expensive resource in Phase 1 was not computation but the engineering of candidate generation and false-positive filtering upstream of any model.

The pipeline is orchestrated by eight numbered runbooks (environment setup, base dataset, timelines, reviews, generation capacity, page counts, technology extraction, geography), each documenting the exact commands, expected outputs, and caveats for one stage. A hard environment guard — the entry-point script refuses to run outside the `nepa` conda environment — keeps runs reproducible.

The Six Deliverables, Technically

Deliverable 1 — The Decarbonization Landscape

The question: how many decarbonization projects are in the dataset, and how do they break down by technology, lead agency, and location?

Methodologically this is the simplest deliverable — no machine learning, no LLM. It unnests NEPATEC's project-type tags into long form and counts projects by technology, department, state, and county, with choropleth maps built from Census TIGER/Line shapefiles using Jenks natural-breaks classification. Its value is descriptive: it establishes the universe every later deliverable draws on, and it surfaces the dataset's concentration — DOE leads 81 percent of decarbonization projects, and the top locations (Aiken County, South Carolina; Boundary County, Idaho) reflect DOE facility sites rather than commercial energy development.

Its main technical contribution is a diagnosis of the location data. County information is present for roughly 45 percent of projects in the frozen universe (about 44 percent of CEs, 81 percent of EAs, and 65 percent of EISs), and the gap has two distinct causes. For EA and EIS projects, counties are usually recoverable: nearly all of the missing cases have usable coordinates that could be reverse-geocoded against county boundaries — a documented, low-effort fix that was scoped but not implemented in Phase 1. For CE projects — 94 percent of the dataset — locations are frequently recorded only as Public Land Survey System legal descriptions ("T. 25 S., R. 22 E., Section 16") with no coordinates, which cannot be geocoded reliably. This asymmetry is a dataset problem, not an extraction problem, and it reappears in the recommendations.

Full report: Deliverable 1.

Deliverable 2 — Programmatic and Tiered Reviews

The question: how common are programmatic reviews and the tiered reviews that build on them, and are tiered reviews faster?

The method is a two-tier cascade. Titles are checked first: terms like "programmatic," "PEIS," or "PEA" in a document title classify the project immediately at high confidence with no document reading. Projects not resolved by title get a regex scan over the first 60 pages of their main documents, with patterns for programmatic language and for "tiered from" references that identify the parent review. A false-positive filter handles the traps discovered during development — chiefly that solar and wind documents routinely cite "EPA Tier 4" diesel-engine standards for construction equipment, which has nothing to do with NEPA tiering, along with road classifications and tiered pricing language. An LLM third tier exists in the code but was not needed: every published classification derives from the deterministic tiers.

The published results classify 161 of 1,326 EA/EIS projects (12.1 percent) as non-standard — 128 programmatic, 33 tiered. On speed, the deliverable's published finding is that tiered reviews do not systematically complete faster: tiered EAs with complete timelines (n=20, median 734 days) ran slower than standard EAs (n=313, median 421 days). Tiered EISs did post the fastest EIS median (593 days versus 914 programmatic and 1,087 standard), but on only seven observations — a figure the deliverable itself flags as too small to interpret reliably.

Validation was structural plus human: automated consistency checks over the classification logic, and a review packet listing every non-standard project with its evidence text, prepared for client validation. The documented limitations are specific: tiering language appearing after page 60 is missed, file names are not searched (at least one project says "programmatic" only in its file name), and the counts are best read as a conservative lower bound on tiering.

Full report: Deliverable 2.

Deliverable 3 — Review Types, Timelines, and Generation Capacity

The question: how do CE, EA, and EIS usage compare — in project counts, in review duration, and in the generation capacity they cover?

This deliverable carries Phase 1's two hardest extractions. The first is timelines. Ten compiled regex patterns detect dates in every format the corpus uses (spelled-out, numeric, ISO, digital-signature syntax), each captured with 150 characters of context. A classifier then labels each date's role — decision, initiation, review activity, or other — trained by weak supervision: strong, medium, and weak signal patterns ("record of decision" and digital-signature syntax are strong decision cues; "notice of intent" and "scoping meeting" are strong initiation cues) auto-label a training set, corrected by hand where needed, with EA/EIS examples oversampled. A BERT classifier scores CE dates; EA and EIS projects — fewer, longer, and messier — get LLM adjudication over their top-ranked candidates. Final selection combines pattern strength, classifier confidence, document-type priority (RODs and FONSIs outrank other documents), and a chronology rule that discards implausibly early dates. Complete-timeline coverage lands at 62 percent for EAs, 48 percent for EISs, and 30 percent for CEs; median durations are roughly one month for CEs, 14 months for EAs, and 34 months for EISs.

The second is generation capacity, which NEPATEC does not record at all. A metadata-first regex pass (titles, then descriptions) precedes three document passes of increasing depth, with four regex template families covering the ways documents state capacity ("80-MW solar facility," "up to 250 MW," ranges, spelled-out units), longest-first unit matching so megawatt-hours are never truncated to megawatts, and power and energy tracked in separate fields because they measure different things. Claude Haiku adjudicates only when two or more distinct candidates survive. Coverage tracks document length: 81 percent of EISs yield a capacity figure but only 8 percent of CEs, whose one-page determinations rarely state one. Spot-check validation found no errors among verifiable sampled extractions (11 of 11, with LLM-adjudicated samples at roughly 90 percent) — and a documented forensic review of every implausibly large CE solar value traced each one to a real pattern (transmission actions citing an upstream farm's capacity, minor modifications citing whole-facility size) rather than to regex failure.

Full report: Deliverable 3.

Deliverable 4 — Geography and Interagency Collaboration

The question: which projects span multiple states or involve multiple federal departments, and which departments sit at the center of the collaboration network?

Multi-state status comes from project metadata. Multi-department status is measured two ways, deliberately: a strict definition from metadata alone (multiple lead agencies recorded), and an expanded definition that adds high-confidence co-agency mentions extracted from EA/EIS text. The text extraction is intentionally narrow: it captures only six explicit NEPA role labels — lead, joint lead, co-lead, cooperating, participating, and responsible federal agency — plus prose lists of cooperating agencies, normalized against an agency-name alias table. Consultation contacts under other statutes (Endangered Species Act Section 7, Clean Water Act Section 404) are excluded by design, a boundary the analysis verified rather than assumed.

The published results: 841 multi-state projects and 301 multi-department projects, the latter 89 percent EISs. Department centrality is summarized by a bridge score — unique partner departments weighted by log collaborative ties — under which Interior (64.2) narrowly leads DOD (62.8) and DOE (50.3). One instructive artifact of the department-level coding is documented in the supporting analysis: splitting the Army Corps of Engineers out of DOD inflates apparent collaboration, because two-thirds of DOD's collaborative projects involve both the Corps and another DOD component — so the split double-counts single projects as two "collaborators." The published figures keep departments intact and treat the Corps breakout as a labeled sensitivity view.

Full report: Deliverable 4.

Deliverable 5 — Document Length and the FRA Page Limits

The question: how long are EAs and EISs, and did lengths change after the Fiscal Responsibility Act (enacted June 3, 2023) imposed page limits — 75 pages for EAs, 150 for EISs (300 if extraordinarily complex)?

The technical problem is that the statutory limits count *regulatory* pages — 500 words per page, excluding citations, maps, and appendices — while the dataset records only raw PDF page counts. The pipeline estimates regulatory pages two ways. Where an agency filed a "without appendices" document variant (detected by filename pattern), that variant's page count is used directly. Otherwise, a DuckDB scan detects the appendix boundary from page-heading text ("APPENDIX A" and variants in the first characters of a sparse page, with a guard against table-of-contents false positives), counts words on the body pages before that boundary, and converts words to regulatory pages at the statutory 500-words-per-page rate. Projects with no OCR-extractable text fall back to raw page counts, flagged as such.

The sample is deliberately strict: final EAs and final EISs only (drafts excluded), decarbonization projects with complete timelines, FRA status assigned by decision date — 627 projects, of which 75 are post-FRA (33 EAs, 42 EISs). Mean regulatory page counts fell about 22 percent for EAs and 25 percent for EISs post-FRA, with roughly seven in ten post-FRA EAs meeting their limit but only about a quarter of EISs meeting theirs. Two caveats bound the reading. The post-FRA window is barely two years and page counts were already trending down before enactment, so the comparison is descriptive rather than causal; and the appendix-detection heuristic has known failure modes (in roughly 5 percent of cases, table-dense documents inflate word counts enough that estimated regulatory pages exceed raw pages).

Full report: Deliverable 5.

Deliverable 6 — Transmission, Geothermal, and Pipelines

The question: for three technologies central to decarbonization — transmission lines, geothermal energy, and pipelines — what do their reviews look like in detail: line lengths, project phases, new construction versus maintenance?

Transmission is the deepest of the three. A four-gate funnel narrows 7,697 transmission-tagged projects to genuinely new construction with measurable length: build-language regex over titles and descriptions (1,741), a maintenance exclusion (1,419), and a one-mile length floor, yielding 263

projects with extracted line lengths (median 11.1 miles). Length extraction is candidate-based: sentences mentioning length-bearing terms yield mileage candidates, filtered for known false positives — geographic distances ("26 miles north of Helena"), right-of-way widths in feet, and milepost references — then resolved by a rule cascade (unique match, take-maximum, sum-of-segments) with Claude Haiku adjudicating genuine multi-candidate conflicts. A dated validation memo reviews the adjudications case by case, records both correct and incorrect outcomes with project identifiers, and concludes that remaining failures trace to upstream candidate filtering rather than model choice.

Geothermal projects are classified by development phase (exploration through utilization) in two stages: keyword rules over titles, descriptions, and document names first; a fine-tuned SciBERT classifier for the remainder. NEPA actions are then rolled up into inferred projects by normalized title matching — 873 actions resolve to 753 projects — because a single geothermal development generates separate NEPA reviews at each phase. The documented caveat is that phase combinations derived from the classifier (as opposed to keyword rules) cannot be decomposed into their constituent phases, which limits one published figure.

Pipelines are identified by tag (5,324 projects), narrowed to new construction by build-language and maintenance gates (899, 17 percent), and split into six technology groups. Carbon dioxide and hydrogen pipelines — nearly all of which pass the new-build gate — are reclassified as decarbonization infrastructure, overriding NEPATEC's fossil-fuel tag for those two groups.

Full report: Deliverable 6.

Data Products & Validation

Phase 1's outputs are parquet files under `phase1/data/analysis/`, one per extraction, all keyed by project and merged into the frozen `projects_combined.parquet` that Phase 2 later builds on. The principal products: the combined project table (61,881 rows × 97 columns); timeline tables (BERT-scored CE dates, LLM-adjudicated EA and EIS dates, and a targeted re-adjudication for the programmatic/tiered subset); the review-type classification; generation-capacity tables with full audit trails; regulatory page counts; the co-agency and collaboration tables; and Federal Register Notice of Intent matches. Document text itself lives in page-level parquet stores per review type, read through DuckDB.

Validation in Phase 1 was manual and targeted rather than statistical. Each extraction shipped with the checks its risk profile warranted: spot-check samples against source documents (generation capacity: all verifiable sampled values correct; LLM-adjudicated samples roughly 90 percent); case-level forensic reviews of anomalies (every implausibly large CE solar capacity traced to its cause; transmission length adjudications reviewed by project ID in a dated memo); structural consistency checks (review-type logic, visual reconciliation of published tables); and client validation packets listing every non-standard classification with its supporting evidence text. Audit-trail columns on every extracted field — method, alternatives considered, model reasoning where applicable — make any individual value traceable to its source page.

What Phase 1 did not have is equally important for a replicator: no frozen held-out gold sets, no formal precision/recall scores for most extractions, and no automated QA assertion suites. Those disciplines were introduced in Phase 2, whose technical report describes them; a reader comparing the two phases should weight Phase 1 numbers accordingly.

Limitations

Four limitations frame everything above.

Coverage asymmetry. The corpus is comprehensive for EISs governmentwide since 2012, but EA and CE records are substantially complete only for DOE and BLM. Coverage is thin for the Army Corps of Engineers, the Forest Service, the Bureau of Ocean Energy Management, and FERC — agencies whose reviews matter for waterways, forests, offshore energy, and gas infrastructure respectively. Every EA and CE statistic describes this particular corpus, not the federal government.

Extraction is measurement, with error. PNNL reports at least 90 percent accuracy for its metadata enrichment, and Phase 1's own extractions carry the spot-check results described above — but natural-language extraction has irreducible measurement error, and findings inherit it. The timeline extraction is the binding constraint: dates were recovered from document text because no federal system records them, an approach the Phase 1 report itself describes as "an inefficient and unwieldy process that is difficult to validate." Complete-timeline coverage of 30–62 percent depending on review type means duration statistics describe the dated subset.

Documents bound what can be known. CE determinations are one to two pages and often simply do not state an initiation date, a capacity, or a precise location; FONSIs and RODs are frequently filed as separate documents that NEPATEC does not link to their parent reviews, which is the largest single cause of missing decision dates. These are facts about federal record-keeping, not correctable extraction defects.

Counts are not quality judgments. A count of categorical exclusions says nothing about whether each CE determination was appropriate, and many CEs cover administrative actions and research funding rather than project deployment — distinctions the dataset does not capture. Similarly, the post-FRA page-count comparison rests on a two-year window and a declining pre-trend, and is presented as descriptive, not causal.

Recommendations

The recommendations below concern the dataset and the records that feed it — the changes that would have made Phase 1's hardest extractions unnecessary. They restate, with the benefit of a completed build, the data-infrastructure recommendations of the Phase 1 report; its policy recommendations are out of scope here.

Publish categorical-exclusion determinations across all agencies. The report's first recommendation is the foundation for everything else: the federal government "should invest in increased data management and transparency for NEPA reviews, including by publishing categorical exclusion determinations across all agencies to allow for broad-based crosscutting analysis." CEs are 91 percent of energy-related NEPA activity in this corpus; as long as most agencies do not publish them, no analysis of the government's dominant review pathway can claim to be governmentwide.

Record review dates as structured metadata. Every date in Phase 1's timeline analysis was mined from prose. The report's assessment bears repeating: "One invaluable data point would be consistent, easily extractable project timeline data. In its current form, all timeline data were extracted directly from document text: this is an inefficient and unwieldy process that is difficult to validate. Future iterations of the NEPATEC dataset would benefit greatly from including this

as metadata." Requiring a summary of key dates in every NEPA document — or better, recording initiation and decision dates in agency systems — would replace Phase 1's most complex pipeline with a lookup.

Link decision records to their reviews. FONSIs and RODs exist as signed, dated documents, but NEPATEC frequently carries them unlinked to the EA or EIS they conclude. That missing linkage — a document-management practice, not a new data requirement — is the largest single cause of missing decision dates in the timeline analysis.

Standardize section structure, terminology, and file names. The same content appears under different names across agencies — the visual-impact section alone is "Visual Resources" (BLM), "Aesthetics" (DOE, NRC), "Scenic Resources" (Forest Service), or "Viewshed" elsewhere — and file-naming conventions vary as freely. Standard section headers and file names would, in the report's words, let "a researcher reliably locate the same section across every EIS," and would sharply reduce the custom engineering every analysis currently requires.

Record locations and capacities as data. CE locations recorded only as Public Land Survey System legal descriptions leave most of the dataset un-mappable; coordinates or county fields would fix this at the source. Generation capacity — central to any energy-policy use of the dataset — appears nowhere in the metadata and had to be extracted with 8 percent coverage for CEs. Both are single fields an agency knows at determination time.

Invest in the data infrastructure that already has vehicles. Phase 1's conclusion stands: existing efforts such as PermitAI, CEQ's NEPA Data and Technology Standard, and the proposed ePermit Act "would all be valuable steps to improve data quality and subsequent research findings — and significant additional investments in federal data infrastructure are still needed." Phase 1 bears out both halves of that sentence: what analysis is possible when documents are at least collected in one place, and how much effort remains when nothing beyond the documents is structured.

Replication Guide

Phase 1 is frozen at the git tag `freeze/v1.0` (`git checkout freeze/v1.0` reproduces the exact published state; the alias `freeze/phase1_v1.0` marks the same freeze for Phase 2's dependency pin). The repository's `phase1/README.md` is the entry point: it maps the directory structure, lists the key facts, and orders the runbooks.

The environment is a single conda environment, `nepa` (Python 3.12), defined by `environment.yml` at the repository root, with the R package set documented alongside it. Extraction requires two credentials: a Hugging Face login to download NEPATEC 2.0, and an `ANTHROPIC_API_KEY` for the adjudication steps (estimated total Phase 1 API cost was on the order of five dollars). The eight runbooks under `phase1/runbooks/` — environment, base dataset, timelines, reviews, generation capacity, page counts, technology, geography — document each stage's exact commands, expected outputs, and caveats, in dependency order.

One disclosure matters for a fresh clone: by design, the repository commits code, documentation, figures, and the final `projects_combined.parquet`, but not intermediate parquets, page-text stores, or model checkpoints. Reproducing the pipeline from scratch therefore starts with runbook 01, which downloads NEPATEC 2.0 from Hugging Face and rebuilds the local data layer; the downstream runbooks then regenerate every intermediate product. Analysis and figures, by contrast,

can be reproduced directly from the committed combined table for most deliverables. The published reports render with Quarto from `phase1/reports/`, each ending with the commands that regenerate its numbers.