Technical Report & Recommendations

Phase 2: Extended NEPA Analysis of the NEPATEC 2.0 Dataset

July 31, 2026

Table of contents

Download PDF version

Executive Summary

This report is a technical companion to the six Phase 2 deliverable reports. It explains what was built, how the pipelines work, how well their methods were validated, 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, written for policymakers and practitioners who want to understand or reproduce analysis of the NEPATEC 2.0 dataset.

Phase 2 built six analytical layers on a universe of 31,508 federal energy reviews (20,725 decarbonization, 10,783 fossil fuel) drawn from NEPATEC 2.0: a classification of why NEPA was triggered for every decarbonization project (99.6% resolved); an extraction of significance determinations from

Findings of No Significant Impact and Environmental Impact Statements, validated against a held-out gold standard; a decarbonization-versus-fossil comparison of review patterns and visual-impact treatment; a timeline database of initiation and decision dates covering the full 61,881-project NEPATEC inventory; an analysis of categorical-exclusion surges after ARRA, the BIL, and the IRA, attributed through in-document law citations; and a systematic screen of FONSI patterns for categorical-exclusion opportunities, verified against the eCFR.

Three facts about how Phase 2 was built explain most of what follows. First, none of these records exist as structured data anywhere in the federal system — every classification, determination, and date had to be reconstructed from document text, agency registers, and public APIs. Second, the pipelines are tiered: cheap, deterministic, fully reproducible methods do the great majority of the work, and large language models are reserved for the small residual of genuinely ambiguous cases — which is why total LLM spend for the phase was on the order of $150–175, and why every model-dependent step has a committed replay record that reproduces the published results without an API call. Third, validation was built alongside each pipeline rather than added at the end: frozen train/test splits, held-out gold sets, automated QA assertions, and explicit statements of what has *not* been validated (most importantly, end-to-end date accuracy in the timeline database).

The report closes with recommendations addressed to the dataset and the process that produces it — structured timeline metadata, governmentwide publication of categorical-exclusion determinations, standardized terminology, and continuous ingestion — the changes that would most reduce the distance between the questions policymakers ask and the data available to answer them.

The Dataset

Phase 2 is built on NEPATEC 2.0, a text corpus of federal NEPA review documents assembled by Pacific Northwest National Laboratory's PermitAI project and published on Hugging Face. NEPATEC 2.0 aggregates Categorical Exclusion (CE), Environmental Assessment (EA), and Environmental Impact Statement (EIS) documents across federal agencies, with full document text broken into page-level records suitable for large-scale text analysis. It is the single external data source underlying every deliverable in this report; nothing in Phase 2 was built from primary agency filings that fall outside this corpus.

From that corpus, Phase 2 defines its analysis universe by an energy-type filter applied to project metadata: 20,725 projects classified as decarbonization and 10,783 classified as fossil fuel, for a combined base universe of 31,508 energy projects. This is the denominator for every cross-cutting deliverable (D1, D3, D5, D6); D2 and D4 further restrict or extend this universe for their own methodological reasons, documented in their respective sections.

Document composition differs sharply by energy type and review depth. Categorical Exclusions dominate both groups — the review type used when an agency determines, without preparing an EA or EIS, that a category of actions does not individually or cumulatively have a significant environmental effect. Environmental Assessments and Environmental Impact Statements are comparatively rare but carry the bulk of the narrative text that later deliverables mine for significance findings and visual-impact language.

Energy group	CE	EA	EIS	Total
Decarbonization	19,399	573	753	20,725

Energy group	CE	EA	EIS	Total
Fossil Fuel	9,191	969	623	10,783
Total	**28,590**	**1,542**	**1,376**	**31,508**

Categorical Exclusions account for roughly 91 percent of the combined universe; Environmental Assessments and Environmental Impact Statements together account for the remaining 9 percent, split almost evenly. This imbalance shapes methodology throughout the report: analyses that depend on narrative environmental-consequences text (significance determinations, visual-impact framing) are necessarily restricted to the EA/EIS minority, while analyses of review-type selection and categorical-exclusion citation patterns can draw on the full universe.

NEPATEC 2.0 records what agencies published, but it does not natively contain authoritative decision dates, funding mechanisms, or a machine-readable catalog of existing categorical exclusions. Phase 2 closes these gaps with four external sources, each addressing a specific extraction weakness.

The BLM National NEPA Register (eplanning.blm.gov) supplies authoritative initiation and decision dates for Bureau of Land Management projects, which make up roughly 41 percent of the combined EA/EIS corpus. Matching NEPATEC case numbers directly to the register's own project records produces dates that do not depend on parsing prose, and they are treated as the most reliable evidence available wherever they exist.

The DOE ePlanning register and an energy.gov Categorical Exclusion listing crawl serve the same purpose for Department of Energy projects. The register/project-page pipeline recovers FONSI, ROD, and NOI dates by matching DOE's own document numbering scheme; the CX crawl separately captures roughly 35,000 categorical-exclusion determination records, since CE determination date is the only NEPA date available for that review type and DOE does not expose it through the standard register.

Federal Register Notice of Intent and Notice of Availability matching adds initiation and end-of-process dates for projects where NEPATEC text cites a specific Federal Register document number. This source replaces an earlier keyword-search approach: matching on the project's own cited document number, rather than searching by title, avoids the coverage loss and false-positive risk of text-similarity search across a large public register.

Finally, a compiled catalog of existing categorical exclusions — roughly 2,105 across 78 agency units, drawn from the eCFR and agency NEPA procedures — gives Phase 2 a reference standard against which project-level CE citations and observed FONSI patterns can be compared: for example, whether a recurring pattern of mitigated findings resembles an already-codified exclusion category.

A replicator starting from scratch begins with Phase 1's frozen output, `projects_combined.parquet` (fixed at the `freeze/phase1_v1.0` tag), which is treated as read-only. Phase 2's pipeline entry point, `extract_data.py`, rebuilds an enriched Phase 2 version of this project-level table — merging in Federal Register enrichment and other Phase 2 additions — and writes it, along with all downstream outputs, to `phase2/data/`. The Phase 1 source file is never modified by any Phase 2 script; every deliverable's data flow starts from this single combined-projects table plus the CE/EA/EIS page-level parquets.

Pipeline Architecture

Two design choices shape every pipeline in Phase 2: the shift to a DuckDB-based data layer, and a tiered-cascade extraction pattern that reserves expensive models for the cases nothing cheaper can resolve.

Phase 1's pipeline was pandas-based. Phase 2 moved to DuckDB for a simple reason of scale: the full NEPATEC page corpus runs to several million pages across the CE, EA, and EIS parquets, too large to load into memory with pandas. DuckDB's `read_parquet()` with predicate pushdown lets every stage — Federal Register document-number scanning, purpose-and-need section extraction, visual-impact section identification — scan the full corpus on disk rather than in memory, and this pattern is used identically whether the task is a one-time metadata join or a full-text regex scan across all pages.

The second recurring pattern is a tiered cascade: cheap, deterministic methods run first, and only the residual cases that survive every deterministic and lightweight-model check are escalated to a large language model. D1's classification of why NEPA was triggered for each decarbonization project is the clearest example. It runs through six ordered tiers — hand-labeled examples, agency-metadata rules, title and description keyword matching, document title scanning, purpose-and-need section extraction, and a fine-tuned SetFit classifier for the hardest agency-specific subset — before an entailment-based model adjudicates any remaining ambiguity. Only the cases that fail every one of those checks are sent to a large language model. In the confirmed run, that final tier processed 501 of 20,725 projects, about 2.4 percent of the universe, and resolved 418 of them. D4's date-selection pipeline follows the same logic in a different domain: register-sourced metadata dates (BLM, DOE, Federal Register) are treated as the most authoritative signal and are never overridden; document-text extraction with a learned classifier and ranker comes next; and full language-model adjudication is reserved for the roughly 11,200 cases (out of 61,881 projects) where a date is genuinely missing and a plausible candidate exists to resolve it.

The same split applies once a case reaches a model: deterministic parsing recovers exact facts wherever facts are structured — a register date field, a CFR citation, a dollar amount in a funding table — while a model is asked to make a judgment call only where the underlying question is genuinely interpretive, such as whether a passage constitutes a significant impact finding or which of several plausible trigger categories best fits ambiguous project language. Where a deliverable needs many thousands of these judgment calls at once, it uses the Anthropic Batch API, which processes requests asynchronously at a 50 percent price discount; where a deliverable's queue is smaller or needs a fast turnaround, it uses direct concurrent calls at standard pricing instead. Every language-model call across every deliverable is made with a named, versioned model and logged with token counts and cost, which is what makes the totals below auditable rather than estimated.

Validation follows the same tiered logic in reverse: every model-based tier is checked against a held-out, hand-labeled gold set before it is trusted, and nondeterministic tiers (SetFit and entailment inference, and every LLM call) are backed by a committed replay record, so that a reviewer can reproduce the exact published classification without re-calling any API. The Data Products & Validation section describes these checks in more detail.

Every call is logged, so the phase's LLM spend can be reported pass by pass:

Deliverable / pass	Model	Volume	Cost
D1 Tier 5 trigger fallback	claude-haiku-4-5	501 calls	~$1.80
D2 FONSI significance (full run)	claude-sonnet-5 (Batch API)	3,478 windows	~$15
D2 EIS significance (full run)	claude-sonnet-5 (Batch API)	21,852 windows	~$110–115
D4 timeline adjudication (cumulative)	claude-haiku-4-5	11,264 calls	$18.28
D6 condition re-tag	claude-haiku-4-5	~11,246 unique calls	~$4.23
D6 FONSI enrichment (39-field extraction)	claude-sonnet-4-6	451 calls	not published as a single total*

*Token volumes (roughly 5,000 input / 1,700 output tokens per call for the extraction stage, an order of magnitude less for the cheaper classification re-ask) are documented, but the architecture notes do not report a consolidated dollar figure for this pass; based on its token profile it is comparable in magnitude to the re-tag pass above, not a dominant cost.

Summing the confirmed figures alone puts total measured LLM spend for the phase at roughly $150–155, and including the unpublished but similarly-scaled D6 enrichment cost keeps the full phase within roughly $150–175. That is a small total for six deliverables covering tens of thousands of classified projects and hundreds of thousands of extracted candidate dates and significance determinations.

The Six Deliverables, Technically

Deliverable 1 — Why NEPA Was Triggered

The question: for each of the 20,725 decarbonization projects in scope, which specific federal nexus triggered NEPA review — funding, land, permitting, or direct federal action among them — and how confident is each answer?

The method is a six-stage cascade that runs increasingly expensive tools only on what earlier stages could not resolve. It opens with 1,473 hand-labeled examples that seed the downstream classifier and cannot be overwritten by anything that follows. Agency-metadata heuristics then map lead agencies with unambiguous jurisdiction — FERC and FAA to federal permit, the Power Marketing Administrations and TVA to their own class — directly to a trigger. A regex pass over project titles and descriptions, extended by an automated scan of document titles and purpose-and-need sections, resolves most of what remains. The largest ambiguous population, DOE-led categorical exclusions, is then classified by a small fine-tuned text classifier that must clear two thresholds at once, a minimum confidence and a minimum margin over the runner-up class; those thresholds were lowered mid-project after the original, stricter pair yielded zero classifications despite hundreds of inference passes. Cases that still resist classification move to an automated adjudication step that tests each

candidate trigger as a natural-language hypothesis against retrieved document text, accepting an answer only when score, margin, and a supporting-evidence check all pass together. The final residue, about two percent of the universe, is reviewed by claude-haiku-4-5 at zero temperature, which returns a structured classification and confidence level.

Projects are sorted into eight mutually exclusive classes under a strict priority order (federal program outranks direct federal action, which outranks the Power Marketing Administrations and TVA, and so on down to federal funding and, last, unknown), because many projects show more than one plausible nexus and a consistent tie-breaking rule is needed to assign one primary class per project. Every secondary nexus detected is retained separately so combinations can still be studied. The pipeline resolved 20,642 of 20,725 projects, 99.6 percent, leaving 83 unknown.

A pre-publication check found a real error: the final review stage's own internal ranking of a project's candidate classes disagreed with the priority order on 88 cases, and all 88 were reconciled and corrected. Because that stage is not perfectly repeatable run to run, exact replication is guaranteed by a committed record of its verdicts that can be replayed deterministically without an API call, rather than by re-running the review itself. The two upstream classifiers that use trained models rather than fixed rules also show a small amount of run-to-run variation near their decision thresholds, measured at under a tenth of a percent of the universe.

The main limitation is that the unresolved 83 are not a random cross-section: 51 of them, about six in ten, are led by the Department of Energy, so DOE-led review remains the weakest-covered slice of an otherwise well-resolved dataset. Fossil fuel projects are out of scope for this deliverable; trigger classification was not attempted for them.

Full report: Deliverable 1. Method and coverage detail: classification scheme and coverage and limitations.

Deliverable 2 — Determinations of Significance

The question: for decarbonization and fossil fuel projects reviewed under an Environmental Assessment or Environmental Impact Statement, what significance determination did the reviewing agency reach across twelve resource areas, and how mitigation-dependent was each finding? Categorical Exclusions are excluded by design; they do not contain the resource-by-resource analysis this question needs.

The method runs two parallel tracks, one for Findings of No Significant Impact under Environmental Assessments and one for Environmental Impact Statements, staged so the roughly nine-times-larger and costlier Statement track was only launched after the first track's results and accuracy numbers were reviewed. Both tracks share the same steps. A deterministic candidate generator first locates document passages likely to contain a significance conclusion, findings and conditions sections for the smaller track, environmental-consequences chapters for the larger one. Each passage is then read in full by claude-sonnet-5, processed through the API's lower-cost asynchronous batch mode, with a prompt that returns a list of determinations rather than a single one, since a passage commonly reaches conclusions on three or more resources at once; an earlier one-determination-per-passage design had captured only about a third of true findings. Passage-length caps were raised to 16,000 characters for the smaller track and 24,000 for the larger one so whole multi-page chapters are read intact.

Coverage narrows at each stage. Of 452 corpus Finding projects, 427 sit within the primary BLM/DOE scope this analysis targets, and of those, 193 projects across 258 documents yield a machine-extractable finding, a limit of the upstream text-location step rather than a sampling choice. The Statement track covers all agencies, since a BLM/DOE-only subset would leave too few projects to analyze; of 753 corpus projects, 506 were ultimately analyzed.

Validation used a gold set hand-labeled by two independent reviewers with a third adjudicating disagreements, 932 labeled instances for the smaller track and 547 for the larger, with 30 percent of each held out purely for accuracy testing. On that held-out share, the model correctly locates a genuine significance finding 97.8 percent of the time for the smaller track and 83.5 percent for the larger, and correctly classifies which determination was reached 80.8 percent and 68.6 percent of the time, respectively. The larger track scores lower throughout because the underlying judgment is harder: the two human reviewers themselves agreed on the right classification only 58 percent of the time, against 68 percent for the smaller track.

The output is one row per significance-bearing determination, keyed so reruns never duplicate a row, and documented field by field in a standalone data dictionary meant to be read alongside the output tables by anyone reproducing this analysis.

Secondary fields trail the primary determination: whether a finding depended on mitigation, and which regulatory threshold it invoked, both score well below the headline accuracy figures. The larger track's pre-launch cost check confirmed only that retrieval was precise, not that it found every relevant passage, so its rates should be read as a well-grounded floor, not a ceiling.

Full report: Deliverable 2. Coverage detail: coverage and limitations; the column-level data dictionary is in the repository at `phase2/notes/deliverable02/data_dictionary.md`.

Deliverable 3 — Review Patterns: Fossil vs. Decarbonization

The question: across the full universe of 31,508 energy projects, 20,725 decarbonization and 10,783 fossil fuel, how does federal environmental review differ between the two — in NEPA review type, in which Categorical Exclusion authorities are cited, and in how visual and scenic impacts are described and rated?

The method runs nine sequential modules over that shared project base. The first builds a common project table with standardized energy-group and technology labels. The second parses every cited Categorical Exclusion in the document set and normalizes four inconsistent citation formats into short, comparable labels: BLM handbook codes, the Interior Department's manual references, Code of Federal Regulations citations, and the standalone statutory exclusion at Section 390 of the Energy Policy Act of 2005. Because the raw citation data is drawn from the entire document corpus rather than the energy project universe, every clean/fossil comparison first joins it back to the 31,508-project base; 29,751 citation rows land on 28,185 of those projects.

Modules three through eight build the visual-impact analysis, restricted by design to Environmental Assessment and Environmental Impact Statement projects, since Categorical Exclusions rarely contain the resource-analysis narrative this method needs. A heading-aware section-identification step scores candidate passages, prioritizing explicit visual or scenic headings, then broader land-use and recreation sections, then a weaker keyword-density fallback; it yields extractable visual text for 1,591 projects. That text feeds three parallel analyses. A domain-specific framing pass applies a negation-aware phrase lexicon, so language such as "no significant adverse visual impact"

is correctly read as non-adverse rather than misclassified by general-purpose sentiment scoring. A four-topic model, fit on term-frequency features with a fixed random seed for reproducibility, resolves the corpus into four stable, interpretable themes: industrial and infrastructure corridors (804 projects), visual resource contrast ratings and solar glare (432), BLM visual resource objectives and landscape management (286), and wind turbine shadow flicker (69). An element-level extractor separately finds explicit BLM visual contrast ratings — form, line, color, texture among them — wherever a document publishes them in a formal ratings table. The ninth module builds a matched comparison of geothermal projects (873) against land-based (8,664) and offshore (211) oil and gas projects, the closest fossil analogue for siting and review patterns.

Checks here are manual and structural: a stratified twenty-project sample read against source documents across energy group, review type, and extraction method, and a standing reconciliation between the section-level and project-level visual outputs that runs on every regeneration to catch join drift before it reaches the report.

One figure needs a narrower reading than the rest: the visual contrast-rating figure is a formal-table subset of only 63 of the 1,591 visual-corpus projects, useful for comparing rated outcomes within that subset but not representative of the broader corpus. More broadly, no visual-impact finding should be read as applying beyond the 1,591 Assessment and Statement projects with extractable text, a small fraction of the full 31,508-project universe.

Full report: Deliverable 3. Coverage detail: coverage and limitations; extended methods notes are in the repository at `phase2/notes/deliverable03/methods_notes.md`.

Deliverable 4 — Review Timelines

D4 answers a basic but data-intensive question: how long does NEPA review take, from initiation to decision, and how has that changed over time and across review types. It is the largest pipeline in Phase 2, spanning all 61,881 projects in the corpus (Categorical Exclusions, Environmental Assessments, and Environmental Impact Statements, across all energy types).

The pipeline layers four sources of evidence, each consulted only when the one before it cannot resolve a project. First, agency register metadata: BLM ePlanning, DOE's ePlanning system, and the energy.gov categorical exclusion listing supply authoritative start and decision dates directly from agency systems, bypassing document text entirely wherever available. Second, for projects the registers do not cover, a five-tier document retrieval strategy pulls candidate evidence from project documents — structured metadata packets, targeted page slices, section-level retrieval by heading, keyword-scored pages, and a final recovery pass — followed by a regex suite that extracts date candidates with their surrounding sentence and a preliminary role label (initiation, decision, or neither). Third, a three-head SetFit classifier (one shared text model with independent heads for initiation, decision, and Final-EIS-publication probability) scores the ambiguous candidates that regex alone cannot resolve; Platt calibration then converts those raw scores into probabilities anchored to a frozen, hand-labeled test set, and a LightGBM ranking model orders candidates within each project. Fourth, a rules-based selection step picks the best initiation/decision pair per project, with review-type-specific logic — for example, an EIS decision searches for a genuine Record of Decision first and falls back to the Final Environmental Impact Statement's publication date only when no Record of Decision exists anywhere in the corpus, since barely a fifth of EIS projects have one on file. Remaining gaps, where candidate evidence exists but the automated steps still cannot confidently resolve it, are sent to a targeted large-language-model adjudication

pass (Claude Haiku), which reviews the top-ranked candidates or reads document text directly; the full run cost $18.28 across 11,264 calls. A final layer allows human-verified dates to override every automated step.

The output is one row per project — 61,881 in total — carrying initiation date, decision date, duration, granularity, and a full evidence trail back to the source document and sentence.

Validation is component-level rather than end-to-end: on a frozen, held-out set of hand-labeled candidate sentences, the classifier achieves an F1 of roughly 0.88 for both the initiation and decision heads; the Final-EIS head is weaker (F1 0.56), reflecting far fewer training examples for that rarer determination. The limitation is that this measures the classifier's accuracy at scoring individual sentences, not the accuracy of the final selected date for a given project end to end; no formal held-out validation exists at that level.

Full report: deliverable04.html. Method notes: date sourcing, known issues, coverage & limitations.

Deliverable 5 — Categorical-Exclusion Spikes After Major Legislation

D5 asks whether categorical exclusion activity rose after the three major clean-energy funding laws of the last two decades — the American Recovery and Reinvestment Act (ARRA), the Bipartisan Infrastructure Law (BIL), and the Inflation Reduction Act (IRA) — and whether that rise can be tied to the legislation itself rather than to the dataset simply growing over time.

The method has two independent legs. The first is a two-stage law-citation scan: a fast DuckDB keyword pre-filter narrows the document corpus (Categorical Exclusions, Environmental Assessments, and Environmental Impact Statements alike) to candidate pages, and a Python regex suite then confirms genuine citations to each law. Ambiguous short names are disambiguated with a surrounding context window — for example, "Recovery Act" only counts as an ARRA citation when nearby language confirms stimulus-era usage, which prevents false matches against the unrelated Resource Conservation and Recovery Act. The second leg normalizes each project's categorical exclusion code to one of three coding schedules — DOE's 10 CFR 1021 Appendix A/B codes, DOI's 516 DM 11 codes, and Energy Policy Act of 2005 §390 — so that shifts in which type of exclusion agencies invoke can be tracked over time, not just shifts in raw volume.

Every categorical exclusion is anchored to a single point in time using Deliverable 4's decision date, falling back to the initiation date when no decision date is available; this coalescing places 95.3% of categorical exclusion projects on the timeline. Because a raw count of exclusions per year is confounded by the corpus's own document-coverage ramp — the dataset is thin before 2009 and incomplete in the most recent two years — the analysis does not rely on aggregate counts alone to make a causal claim. It relies on two checks that are unaffected by that confound: conditioning on agency (the Department of Energy, which administered ARRA's energy funding, shows a sharp rise while the Bureau of Land Management, subject to the identical coverage ramp, stays flat), and citation evidence, which cannot be an artifact of corpus growth since a law cannot be cited in a document before it exists. Within the ARRA spike window, 59.7% of exclusions explicitly cite the law by name, versus a negligible rate outside it.

There is no labeled answer key here. The check is internal consistency: the citation-rate evidence and the agency-conditioned volume evidence point the same direction, and the categorical-exclusion code-mix shift during the ARRA window (a sharp rise in the "energy and water conservation" code,

consistent with weatherization and efficiency stimulus spending) is substantively plausible rather than an artifact of the extraction method. ARRA has no usable pre-law baseline in the corpus, and the BIL and IRA response windows overlap by sixteen months, so date alone cannot cleanly separate a BIL-driven exclusion from an IRA-driven one in that overlap — citation evidence, where it exists, does the separating work instead.

Full report: deliverable05.html. Method notes: coverage & limitations.

Deliverable 6 — Patterns in FONSIs: Categorical-Exclusion Opportunities

D6 asks a forward-looking question: among clean-energy projects that completed a full Environmental Assessment and received a Finding of No Significant Impact, which recurring categories of action look like credible candidates for a new categorical exclusion, an expanded existing one, or adoption of an existing exclusion by an agency that does not currently use it.

The analysis is built on a corpus of 452 clean-energy, Environmental-Assessment-sourced FONSI projects — reading each project's Finding plus its underlying Environmental Assessment, not just the one-page finding itself. A single large-language-model enrichment pass (Claude Sonnet) reads every project once and extracts 39 structured fields — action description, physical scale, siting characteristics, mitigation dependence, and any explicit significance thresholds the document states — using a cached two-stage design that first extracts and then re-classifies the action category cheaply from the cached extraction. Every quoted value is checked back against its cited source text before being trusted. A controlled vocabulary of eleven action verbs (new build, upgrade, maintenance, exploration, assessment, and others) is then assigned to each project, and crossing that verb with the project's technology group produces a fully enumerated grid of 52 observed action categories — every FONSI lands in exactly one cell, with no residual "other" bucket left unaccounted for.

Each grid cell is compared against the roughly 2,105 existing categorical exclusions across the federal government, using text-similarity matching (median similarity across the cell's member projects, to avoid one atypical project skewing the match). A cell earns a "new" verdict when no existing exclusion is a close match, "expand" when a matching exclusion exists but the cell's projects exceed its stated numeric limits, and "adopt" when a matching exclusion exists at one agency but not at the agencies actually doing this work. A weighted, multi-factor score ranks the resulting candidates, and a 2,000-draw sensitivity sweep over the ranking weights confirms which candidates are robustly near the top versus which are sensitive to how the factors are weighted.

Because a text-similarity match is not the same as confirmed legal coverage, every "adopt" or "expand" candidate's top matches were checked against the current Code of Federal Regulations text. That check surfaced a structural limit: of the 120 top candidate matches examined, only 29 are actually codified in the CFR — the rest live in individual agencies' own NEPA procedure documents or point to outdated regulatory citations — so most candidates are capped at "partially" verified rather than fully confirmed pending agency-procedure-level review.

A 25-assertion script checks the pipeline's internal consistency (every project assigned to exactly one grid cell, verdict logic applied correctly, no broken categories), and 96.7% of cited quotes verify against source text. Even a verified categorical-exclusion match is a text-similarity finding rather than a legal determination: a starting point for legal and policy review, not a substitute for it.

Full report: deliverable06.html. Method notes: coverage & limitations, eCFR verification.

Data Products & Validation

Every figure and table in this report traces back to a small set of published analysis files under `phase2/data/analysis/`, each with a single well-defined grain, a UTC run timestamp column, and a documented schema. The table below catalogs the principal file per deliverable — the one an analyst would open first to check a headline number — not every intermediate or diagnostic file the pipelines produce.

File (deliverable)	Grain	Rows	Purpose
`projects_nepa_triggers.parquet` (D1)	one row per decarbonization project	20,725	Primary and secondary NEPA-trigger classification (federal funding, federal land, permit, direct action, etc.)
`significance_determinations.parquet` (D2, EA/FONSI track)	one row per candidate window × resource area × determination	7,250	Significance determinations extracted from Environmental Assessment and Finding of No Significant Impact documents
`significance_determinations_eis.parquet` (D2, EIS track)	one row per candidate section × resource area × determination	59,357	Significance determinations extracted from Environmental Impact Statement documents
`projects_nepa_reviews.parquet` (D3)	one row per decarbonization or fossil fuel project	31,508	Base review table (NEPA review type, agency, geography, technology, trigger) underlying the decarbonization-vs-fossil comparison
`ce_citations.parquet` (D3)	one row per parsed Categorical Exclusion citation	56,681 (29,751 joined to the D3 project universe)	Normalized Categorical Exclusion authority citations by project
`timeline_project_dates.parquet` (D4)	one row per project, full NEPATEC 2.0 inventory	61,881	Selected initiation and decision dates, duration, and coverage/confidence flags for every review

File (deliverable)	Grain	Rows	Purpose
`law_citations.parquet` (D5)	one row per project × law	8,741	Explicit citations to the American Recovery and Reinvestment Act, the Bipartisan Infrastructure Law, the Inflation Reduction Act, and DOE funding authorities
`ce_categories.parquet` (D5)	one row per project × normalized Categorical Exclusion code	74,035	Normalized Categorical Exclusion authority codes underlying the spike and category-mix analysis
`fonsi_enrichment.parquet` (D6)	one row per decarbonization FONSI project	452	Structured extraction of action, scale, siting, mitigation, and citations underlying the FONSI-pattern analysis
`candidate_verdicts.parquet` (D6)	one row per technology-action cell	52	Final new/expand/adopt/already-covered verdict and ranking for each recurring FONSI action category

Row counts above were confirmed directly against the committed parquet files (not estimated) except for the two Deliverable 5 files and the Deliverable 2 EIS track, which are confirmed against each architecture document's Run Results section. Column-level definitions for every field are documented at the source: each `phase2/architecture/deliverables/deliverableNN.md` file has an Output Schema section covering its own parquets, and Deliverable 2 additionally has a standalone column-by-column data dictionary at `phase2/notes/deliverable02/data_dictionary.md` covering all fourteen of its emitted tables, including the gold-set and validation files.

Validation followed a different method for each deliverable, matched to how that deliverable's classification was built — deterministic rule and citation-matching pipelines were checked by construction and spot check, while the deliverables built on learned models or LLM adjudication were checked against held-out, human-reviewed answer keys.

Deliverable	What was validated	Method	Headline score
D1 — NEPA Triggered	Trigger-classification coverage and quality	Full-corpus resolution rate, plus a manual spot check of the LLM fallback tier	99.6% of projects classified (20,642 of 20,725); a 20-row spot check of the 501-project LLM tier found no clear errors
D2 — Significance Determinations	Determination detection and classification accuracy	Held-out gold set: two independent AI labelers double-coded a stratified window sample, a human analyst adjudicated every disagreement, and 30% of windows were held out from tuning	Window-detection F1 0.978 (EA/FONSI track) / 0.835 (EIS track); determination-class macro-F1 0.808 / 0.686
D3 — Review Process Application	Categorical Exclusion citation parsing, review-table construction, visual-section extraction	Deterministic parsing and joins (no learned model), checked by construction, plus a stratified 20-row QA sample of extracted visual-resource sections	No formal accuracy metric; validated by construction and manual spot check
D4 — Timelines	The candidate date classifier that ranks extracted date evidence	Frozen train/test split held out by candidate identifier	F1 0.882 (initiation head) / 0.885 (decision head) on the held-out test split
D5 — CE Spikes	Law-citation and Categorical Exclusion category extraction	Deterministic regex extraction, cross-checked against an agency contrast (DOE vs. BLM) and citation-based attribution	No formal accuracy metric; validated by construction and cross-check, not against a labeled gold set
D6 — FONSI Patterns	Pipeline invariants and LLM-enrichment quote fidelity	Automated QA assertion suite over the full extraction grid, plus verbatim verification of every cited quote against its source span	25 of 25 QA assertions pass; 96.7% of cited quotes verify against source text

The same validation practices run across all six deliverables, including the four without formal F1 scores. Wherever a model or an LLM is in the loop, its train/test split is frozen — assigned

once, stratified, and never re-randomized as new labeled rows are added — so a later relabeling pass cannot silently leak into an evaluation set; Deliverable 4 additionally maintains a permanent do-not-train registry (`frozen_eval_ids.txt`) that its ranker training script hard-fails against if violated. Held-out gold sets are used wherever human judgment is the reference standard (Deliverable 2's determination gold, Deliverable 1's classifier example bank), and are kept in files separate from the production output so they can be re-scored independently. Because several tiers use a nondeterministic LLM call, exact replication does not depend on re-calling the model: Deliverable 1's Tier 5 fallback and Deliverable 4's LLM adjudication step both write a committed record of every call (`tier5_adjudication_record.csv`; `timeline_api_adjudications.parquet`) that a `--from-record` or cache-replay mode can re-apply deterministically at zero cost. Deliverable 6 runs an explicit 25-assertion QA script (`qa_deliverable06.py`) after every pipeline run, and Deliverable 2 writes a run manifest (`significance_run_manifest.parquet`) recording every emitted file's path, row count, content hash, and model/prompt version. Deliverable 4's hand-labeled training inputs — the candidate-level classifier labels and the project-level verified date picks — are treated as primary source data rather than regenerable output, and live under `phase2/training/deliverable04/` (see that directory's README for the frozen-split discipline).

Limitations

Five limitations cut across all six deliverables and should be read as caveats on the dataset as a whole.

Agency coverage asymmetry. Categorical Exclusion and Environmental Assessment records are substantially complete only for DOE and the Bureau of Land Management — the two agencies that dominate NEPATEC 2.0's non-EIS intake — while equivalent records from the Army Corps of Engineers, U.S. Forest Service, Bureau of Ocean Energy Management, and the Federal Energy Regulatory Commission are thin or largely absent. Deliverable 2's EA/FONSI track makes this explicit by restricting its headline denominator to BLM plus the DOE family (94.5% of that corpus); Deliverable 1's trigger classification was scoped to decarbonization projects only, so Deliverable 3's fossil fuel portfolio carries no trigger classification at all. Any EA-level statistic in this report describes Environmental Assessments in this particular corpus — heavily DOE and BLM — not the full federal EA population; EIS coverage, by contrast, comes from a broader, governmentwide crawl.

Timeline fragility. No field in the underlying agency systems reliably records when a review began or concluded, which is why Deliverable 4 reconstructs both dates from document text, agency registers, and, for a residual set of ambiguous cases, targeted LLM adjudication. A Record of Decision — the document that normally marks an EIS's completion — exists in the corpus for only about 18% of EIS projects (732 of 4,130); where none is available, the pipeline falls back to the Final EIS publication date as a month-granularity proxy, and 43% of all EIS decision dates in the published output are this kind of fallback. EA initiation dates are frequently absent from the source documents themselves: an audit of 100 EA projects with a decision date but no initiation date found that 82 had no initiation signal anywhere in their documents.

Field sparsity. Several extracted fields have thin, not full, coverage and should be read as describing a subset rather than a portfolio total. Federal funding dollar amounts are extracted for only about 6% of the 9,210 federal-funding-triggered decarbonization projects (528 projects); project

county is populated for roughly 47% of the full 61,881-project inventory. Distributional statements (typical award size, geographic concentration) are appropriate uses of these fields; portfolio-wide totals are not.

Coverage is not accuracy. "Complete" in the timeline database means both an initiation and a decision date were found — not that either date has been independently verified against source documents. Deliverable 4's individual model components (the candidate classifier, the selection ranker) are validated on held-out splits, but the end-to-end selected date has not been formally validated against a held-out, hand-verified gold set at the project level. Coverage rates measure how often the pipeline found a plausible date, not how often that date is correct.

Ingestion lag. NEPATEC 2.0's most recent one to two years of records are incomplete because agencies' document postings lag real-world project activity, so 2024 and 2025 counts understate true activity across every deliverable that plots volume by year — Deliverable 5's apparent decline in DOE categorical exclusions from 2023 to 2024 (983 vs. 3,152) is a data-ingestion artifact, not a real drop in program activity. Because the Fiscal Responsibility Act's page and time limits took effect on June 3, 2023, this same lag means every post-FRA comparison in this report rests on a still-incomplete window and should be refreshed as NEPATEC's ingestion catches up.

Each deliverable's full coverage-and-limitations discussion, with the underlying numbers and additional deliverable-specific caveats, is published alongside its report: D1, D2, D3, D4, D5, D6.

Recommendations

The recommendations that follow concern the dataset and the process that produces it, not NEPA policy. Everything Phase 2 built — trigger classifications, significance determinations, timelines, categorical-exclusion analysis — was reconstructed from document text because the underlying records do not exist in structured form. Most of the engineering effort documented in this report, and most of its limitations, trace back to that single fact. The changes below would make analyses like these cheaper, faster, and more reliable, whether performed by this team or anyone else.

Improve what agencies record

Record review milestones as structured metadata. The single highest-value change is for agencies to record when reviews begin and end — initiation, key milestones, and decision dates — as structured data rather than prose buried in documents. Every date in the Phase 2 timeline analysis had to be recovered from document text or agency registers; the pipeline that does so spans register crawls, five retrieval tiers, trained classifiers, and model adjudication, and still cannot reach the dates that documents simply never state. The CEQ NEPA and Permitting Data and Technology Standard already contemplates exactly this metadata; agency adoption would remove the hardest problem in this body of work. The Phase 1 report reached the same conclusion: "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." Phase 2's experience confirms it.

Publish categorical-exclusion determinations across all agencies. EIS records are governmentwide, but EA and CE coverage is substantially complete only for DOE and BLM. The agencies where coverage is thin are precisely the ones whose reviews matter for energy infrastructure: the

Army Corps of Engineers (waterways and wetlands), the Forest Service, BOEM (offshore), and FERC (pipelines, hydropower, LNG). Crosscutting analysis of the fastest NEPA pathway is impossible while most agencies do not publish CE determinations at all. As Phase 1 recommended, 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."

Include decision records in the corpus. Only a minority of EISs in the corpus carry a Record of Decision, forcing the timeline pipeline to fall back on Final-EIS publication dates for roughly 43% of EIS decisions. RODs are signed, dated documents that exist; systematically depositing them with the rest of the review file would replace proxy dates with authoritative ones.

Standardize terminology, section structure, and file naming. The same resource area appears under different names across agencies — visual impacts alone surface as "Visual Resources," "Aesthetics," "Scenic Resources," and "Viewshed" — and section headers and file names vary just as freely. A shared resource-area vocabulary and standard section headers would make determinations directly comparable across agencies and dramatically simplify extraction — in the Phase 1 report's words, "so that a researcher could reliably locate the same section across every EIS."

Publish machine-readable CE catalogs. Verifying candidate categorical exclusions against their authorities required an eCFR check; only 29 of the 120 top matches were codified there, with the remainder living in agency NEPA-procedure documents that exist only as PDFs. A maintained, machine-readable catalog of every agency's CEs — citation, text, bounds, and status — would turn that verification from a research task into a lookup.

Extend what the dataset collects

Four additions would extend what the dataset can answer without changing how agencies work. First, OCR reprocessing: roughly 950 projects in the corpus are scanned documents that yield no machine-readable text and are unrecoverable without it. Second, federal funding amounts: dollar figures could be extracted for only about 6% of funding-triggered projects, enough for distributional statements but not portfolio totals. Third, geography: county information is present for roughly half the corpus, yet nearly all of the gap is recoverable by reverse-geocoding coordinates the dataset already holds — a low-effort, high-value fix. Fourth, verified project geometry: linear-versus-sited classification currently rests on a label heuristic; actual footprints would put spatial analysis on solid ground.

Streamline how the dataset is maintained

NEPATEC is a snapshot; the analyses built on it age accordingly. Counts for the most recent two years are visibly depressed by ingestion lag, and any post-Fiscal Responsibility Act question requires a refresh before the answer can be trusted. Continuous ingestion — with versioned, dated snapshots and a manifest of external sources (registers, the Federal Register, the eCFR) — would turn one-time research exercises into a standing capability. The federal vehicles for this investment already exist: PermitAI, the CEQ data standard, the proposed ePermit Act, and the Permitting Council's Permitting Dashboard are all steps in this direction. Phase 1's assessment still holds — these "would all be valuable steps to improve data quality and subsequent research findings — and significant additional investments in federal data infrastructure are still needed."

Additional use cases

The dataset can support more than it currently does, without new collection. The project-to-document-to-page structure makes precedent research practical: practitioners can locate comparable projects by agency, process, technology, geography, and period, and read exactly what the reviewing agency wrote. The categorical-exclusion screening built in Deliverable 6 is designed to be re-run as the corpus grows — with more post-FRA decisions, the same machinery can surface CE candidates for expert review on a recurring basis rather than as a one-off. The duration questions left open in this phase become answerable in one to two years simply by accumulating decisions. And a future question-answering interface over the corpus — one that keeps exact facts in structured tables and constrains a model to produce only cited summaries — could make the corpus usable by practitioners without database skills. None of these require new data; they require only that the dataset continue to exist and be maintained.

Replication Guide

Every analysis in this report can be rebuilt from the public repository. Code is organized by deliverable under `phase2/code/deliverableNN/`, with shared extraction utilities under `phase2/code/extract/` and the register scrapers behind Deliverable 4's date sourcing under `phase2/code/api/` (BLM ePlanning, DOE ePlanning, and the DOE Categorical Exclusion determination listing). Each deliverable's design decisions, script-by-script behavior, confirmed run results, and known issues are documented in `phase2/architecture/deliverables/deliverableNN.md`; a parallel, more operational runbook covering the exact command sequence, arguments, and validation targets for each deliverable lives in `phase2/runbooks/deliverables/deliverableNN.md`. Narrower methods notes, coverage-and-limitations pages, and (for Deliverable 2) a full column-level data dictionary live under `phase2/notes/deliverableNN/`.

The pipeline runs in a dedicated conda environment named `nepa`; Python scripts are invoked as `conda run -n nepa python phase2/code/deliverableNN/script.py`, and R figure and analysis scripts as `Rscript phase2/code/deliverableNN/NN_create_figures.R`. Page-level document text is stored as partitioned parquet files under `phase2/data/processed/{ce,ea,eis}/` and is always read through DuckDB rather than loaded wholesale into memory — every script that touches page text follows this pattern to keep memory use bounded on a multi-gigabyte corpus.

The exact reproduction commands for the numbers in each public report are given in two places, deliberately kept in sync: the Reproduction section at the end of each report (`phase2/reports/deliverableNN.qmd`), and the corresponding runbook (`phase2/runbooks/deliverables/deliv`
which additionally covers validation targets and troubleshooting the reports omit.

Phase 2 depends on a frozen Phase 1 output as its starting universe; that dependency is pinned at the `freeze/phase1_v1.0` git tag rather than a live path, so Phase 2 results do not silently shift if Phase 1 code is later revised. Hand-labeled gold sets and model training labels are treated as primary source data, not regenerable output: Deliverable 4's are under `phase2/training/deliverable04/`, and Deliverable 2's gold sets live under `phase2/data/analysis/deliverable02/gold/`.