

Evolving the Edison Papers

A Human-in-the-Loop AI Platform for Transcription and Indexing

Partha Pediredla

Advisors: Ivan Seskar, Narayan Mandayam, Paul Israel, Lucy Israel

Mentors: Zihao Ding, Xiaotian Zhou

Rutgers University · WINLAB · Thomas A. Edison Papers

The Problem

The Edison Papers Digital Edition hosts
~**154,000** digitized documents.

Browsing works. Metadata is rich.

But almost none of it is transcribed — so none of it is searchable as *text*.

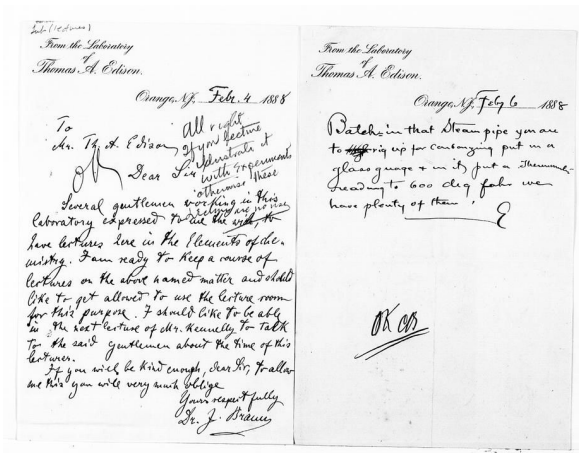
149,619 documents
without a transcription

2.7% transcribed.

4,147 done

The screenshot shows the website interface for 'THE THOMAS A. EDISON PAPERS DIGITAL EDITION'. At the top, there is a navigation bar with links: 'Browse Item Sets', 'Search Items', 'Explore by Topic', 'Indexes and Finding Aids', 'People & Organizations', and 'Edison Papers Website'. Below the navigation bar, the page title 'THE THOMAS A. EDISON PAPERS DIGITAL EDITION' is displayed. Underneath, there is a section titled 'Items' with a 'Template: Base Resource' dropdown. A search bar is visible with 'Created' and 'Ascending' filters, and a 'Sort' button. To the right of the search bar, there is a 'Advanced search' link. Below the search bar, there is a pagination bar showing '1' of '3075' items, with a 'Previous' button and a 'Next' button. The main content area displays four document thumbnails, each with a title and a date. The thumbnails show handwritten text on aged paper. The titles and dates are: '[0184X01BA4].Untitled 1907-11', '[0242X01BA4].Untitled 1928-02', '[0258X01BA4].Untitled 1918-04-29', and '[0259X01BA4].Untitled 1917-00-00'.

Why This Is Hard



One real page from the archive (et0976):

- ▶ Cursive, 1888, two different hands
- ▶ Marginal notes **crammed between lines**
- ▶ Words crossed out and replaced
- ▶ Printed letterhead mixed with handwriting

This is the *normal* case, not the worst case.

How Transcriptions are Measured

Both metrics are **edit distance**, normalized by the reference length:

$$\text{CER} = \frac{S + D + I}{N_{\text{chars}}}$$

$$\text{WER} = \frac{S + D + I}{N_{\text{words}}}$$

S, D, I = substitutions, deletions, insertions.

Note: insertions count on top but not below — so **CER can exceed 1.0**. A model that hallucinates scores 6.11, not 1.0.

Worked example

reference:

dear sir, i enclose the report

prediction:

dear si, i enclose the report now

	Characters	Words
Deleted	1 (r)	0
Inserted	4 (now)	1 (now)
Substituted	0	1 (sir,)
Reference length	30	6
Score	0.167	0.333

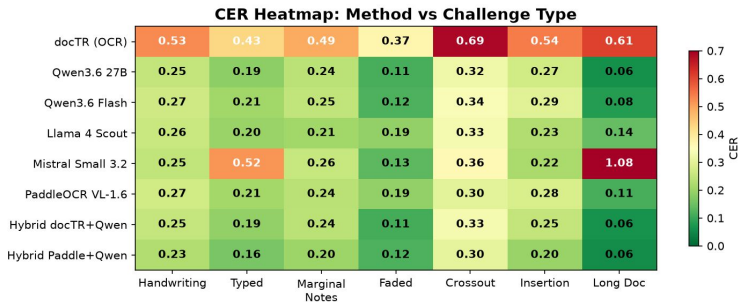
WER is harsher — one wrong character condemns a whole word.

Benchmarked 8 Methods

Method	Type	Docs	CER↓	WER↓	\$/doc
PaddleOCR-VL → Qwen	Hybrid	26	0.198	0.261	0.0076
docTR → Qwen	Hybrid	26	0.220	0.278	0.090
Qwen3.6-27B	VLM	26	0.218	0.276	0.0223
Qwen3.6-Flash	VLM	26	0.235	0.299	0.0069
Llama-4-Scout	VLM	26	0.234	0.324	0.0005
PaddleOCR-VL	OCR-VL	26	0.244	0.320	<i>free</i>
Mistral-Small-3.2	VLM	26	0.452	0.503	0.0003
docTR	OCR	26	0.516	0.823	<i>free</i>

26 documents from **Volume 9**, ground truth from the published book edition, every document hand-labeled with challenge tags.

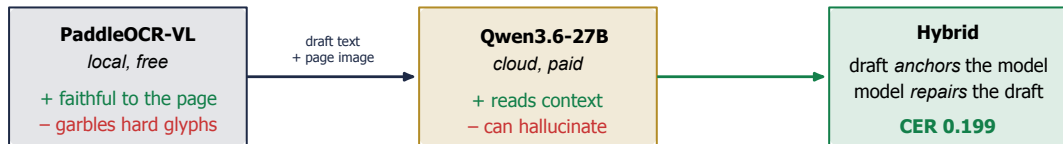
Where Each Method Breaks Down



Green = lower error = better.

- ▶ Crossouts and handwriting are hard for *everything*
- ▶ Classical OCR (top row) simply cannot read 19th-century script
- ▶ The hybrid (bottom row) is the greenest band overall

Why the Hybrid Performs Best in Our Benchmark



- ▶ The model never sees a blank slate — it sees a draft it must justify *against the image*
- ▶ That constraint is what suppresses runaway hallucination
- ▶ **The free half does the reading; the paid half only corrects**

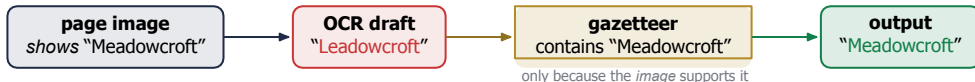
Archive Specific Prompting and Guardrails

1. Structural separation

Tag what the author *didn't* write, inline: [Letterhead] [Stamp] [Marginal note]. Otherwise a researcher can't tell Edison's words from a mailroom stamp.

2. Entity gazetteer

A curated gazetteer repairs "Leadowcroft" → "Meadowcroft" – but only where the page image supports the correction



All three prompts are versioned files; every stored transcription records which version produced it.

3. Conservative splitting

"When uncertain, prefer **continuation** over inventing a split" — a missed boundary is one click to fix; a spurious one is a mess. Duplicate scans are flagged on a *separate* axis.

4. Deterministic fields in code

Titles follow a fixed pattern on the live site, so they're **computed in code**, never generated.

What Was Built: Two Pipelines

Existing Documents (transcription)



New Documents (transcription + indexing)



Shared: the same hybrid engine, the same storage layer (file:// today, s3:// by config), the same review UI.

New on the bottom row: bulk scans have *no document boundaries* and *no metadata* — both have to be produced.

Where Does One Document End?

The screenshot shows the Edison Papers Transcription Review interface. On the left is a sidebar with a 'PIPELINE' section containing 'Upload Scans', 'Draft Pending', and '1 document awaiting drafting'. The main area is titled 'Confirm Document Split' with a subtitle 'Fix any boundary Qwen's proposal got wrong, then name each resulting document.' It features a 'FOLDER ID' field with 'E2231' and an 'Apply Naming' button. A 'Confirm & Draft' button is highlighted with a red box. The interface lists document groups: 'DOCUMENT NAME (GROUP 1)' with 'E2231AAA' and 'DOCUMENT NAME (GROUP 2)' with 'E2231AAB'. Each group contains a list of pages with checkboxes to 'Start a new document'. For Group 1, 'Page 1 - scan_0000.jpg' is checked, while 'Page 2 - scan_0001.jpg' is not. For Group 2, 'Page 3 - scan_0002.jpg' is checked. A 'Cancel' button is also visible.

A scan folder is an ordered pile of images. Nothing records where one letter ends and the next begins.

The model proposes boundaries page by page;
the reviewer confirms or corrects every one
before anything is transcribed.

Scaling: a 70-page folder exceeds the request-size limit, so it is split into overlapping windows dispatched **3 at a time** — **2.26×** faster on a real 71-page batch.

Nothing Counts Until a Human Accepts It

Edison Papers
Transcription Review

Queue

Page 1/1

Reset View

[ET0976AAA] Letter from Dr. J. Stevens to Thomas A. Edison, February 4th, 1888

UNREVIEWED Delete Document

ET0976AAA Edit ID

TITLE (auto-generated from the fields below — add to override)
[ET0976AAA] Letter from Dr. J. Stevens to Thomas A. Edison, February 4th, 1888

DOCUMENT TYPE DATE
Letter 1888-02-04

AUTHOR(S) RECIPIENT(S)
Stevens, Dr. J. Edison, Thomas A.

PLACE MENTIONED
Orange, N. J. Arthur E. Kennelly

SUBJECTS
Electricity; Lectures; Laboratory

TRANSCRIPTION Edit

[Letterhead] From the Laboratory of Thomas A. Edison, Orange, N.J. Feb. 4 1888 To Mr. T. A. Edison Dear Sir I have lectures here in the Elements of Electricity. I am ready to keep a course of lectures on the above named matter and should like to get allowed to use the lecture room for this purpose. I should like to be able in the next lecture of Mr. Kennelly to talk to the said gentlemen about the time of this lectures. If you will be kind enough, Dear Sir, to allow me that you will very much oblige Yours respect fully Dr. J. Stevens [Marginal note] [Letterhead] From the Laboratory of Thomas A. Edison, Orange, N.J. Feb 6 1888 Patch in that Steam pipe you are to rig up for condensing put in a glass gauge + in it put a thermometer reading to 600 deg fah we have plenty of them E

MARGINAL NOTES (editable)
All right from lecture of Elements of Electricity with experiments otherwise these gentlemen working in this laboratory expressed to me the wish

DO A/FEC REVISION/EDIT, UNREVIEW, LOCK/UNLOCK ROWS (this is nonfunctional)

Save Draft Flag Accept Mark Unreviewed

Real output — the same 1888 page from earlier. It read the cursive, tagged the letterhead, pulled the angled marginal note, dated it **1888-02-04**. Cost **\$0.0016**.

Every field editable, raw AI output **never overwritten**, export takes only **accepted** documents.

What It Costs to Run

Measured cost

One single-page letter	\$0.0016
Benchmark avg (multi-page)	\$0.0076
PaddleOCR-VL draft pass	<i>Free</i>

Projected: all 149,619 documents

If mostly single-page	~\$245
If multi-page average	~\$1,420

Why it can't run away

- ▶ A hard **all-time spend cap**, checked before *every* model call — not per run. Reached $\Rightarrow$ drafting stops.
- ▶ Adjustable live in the Settings menu, no restart
- ▶ **Human-paced**: nothing runs in a background loop — a reviewer clicks, a bounded batch runs
- ▶ Split proposal is **resumable** — a mid-batch failure never re-bills finished pages

The bottleneck isn't cost — it's review throughput.

What's Done

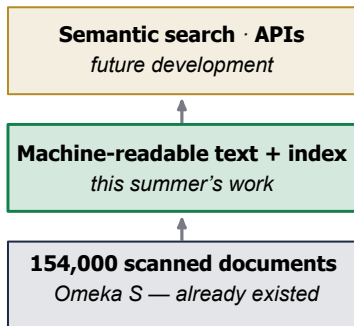
- ✓ Benchmarked 8 methods; picked the hybrid on evidence, not intuition
- ✓ Both pipelines working end to end against the live archive
- ✓ AI split proposal + full metadata indexing
- ✓ Filtered Discovery - the reviewer picks what to transcribe
- ✓ Cost caps, resumability, remote multi-device access
- ✓ **343 automated tests passing**

149,619 documents
have a working path to transcription

\$245 – \$1,420
projected for the entire backlog

Every output provisional
a human accepts it before it counts

For Future Development: Archive → Research Platform



Each layer needs the one beneath it to exist first.

The original goal was to move the site from a *static repository* to a *dynamic research platform*. Semantic search and public APIs were always the destination — but neither is possible over images.

Now that the text and metadata exist:

- ▶ **Semantic search** — embed the transcriptions so a researcher can search by meaning, not just keyword; the indexed places, names and subjects give it structure to filter on
- ▶ **Public APIs** — expose transcriptions and their entity index for external digital-humanities work
- ▶ **Write-back to Omeka** — close the loop so accepted transcriptions land on the live site, not just a CSV
- ▶ **Corpus-scale analysis** — correspondence networks, topic and place trends across 150,000 documents

Questions?