LLMs for Transcription and Metadata Creation

Katie Pierce Meyer

Head of Architectural Collections

Hannah Moutran

Library Specialist

Applications of AI

Karina Sanchez

Scholars Lab Librarian

University of Texas Libraries

Chase Architecture & Planning Library

Josh Conrad

Digital Initiatives Archival Fellow

Devon Murphy

Metadata Analyst

Willem Borkgren

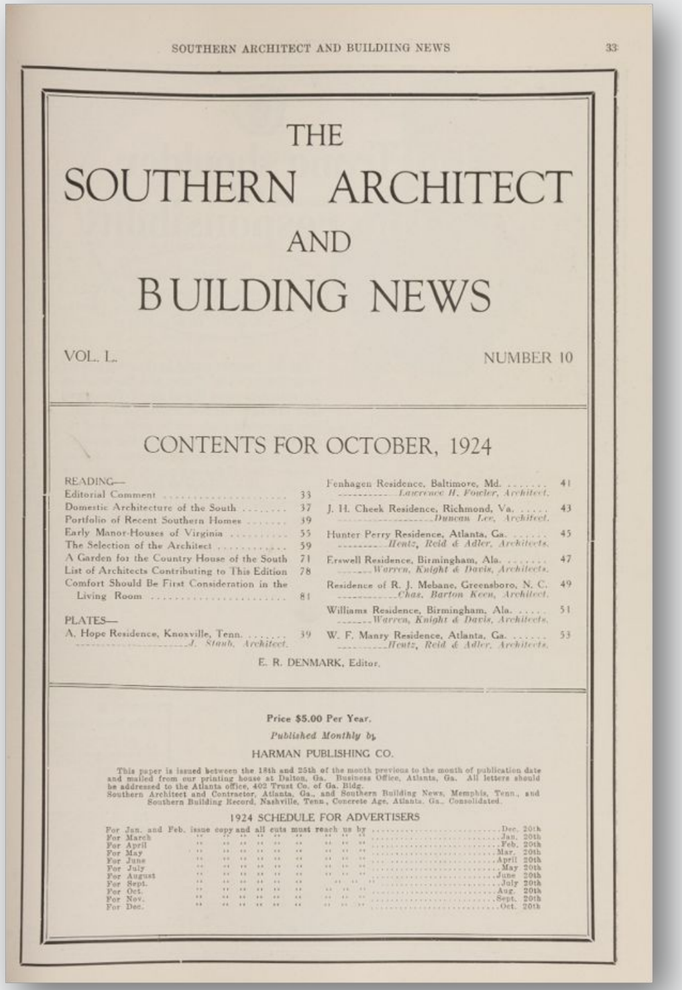
Scholars Lab GRA

Table of Contents

01

At a Glance: Collection & Idea

Southern Architect and Building News Collection (SABN)



SOUTHERN ARCHITECT AND BUILDIING NEWS 33

THE
SOUTHERN ARCHITECT
AND
BUILDING NEWS

VOL. I. NUMBER 10

CONTENTS FOR OCTOBER, 1924

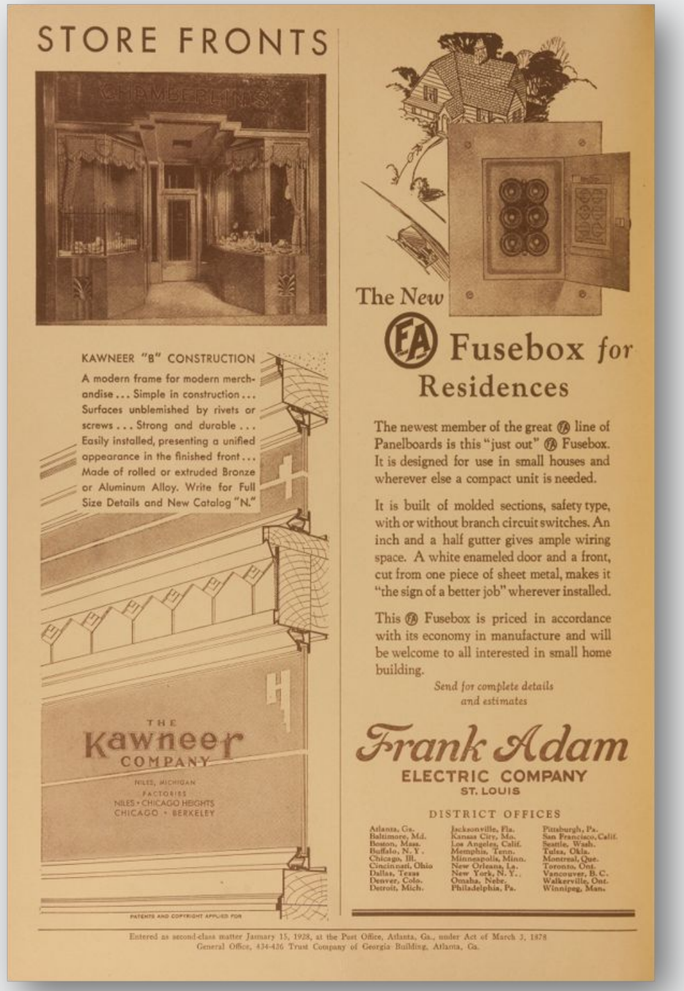
E. R. DENMARK, Editor.

Price $5.00 Per Year.
Published Monthly by
HARMAN PUBLISHING CO.

This paper is issued between the 18th and 25th of the month previous to the month of publication date and mailed from our printing house at Dalton, Ga. Business Office, Atlanta, Ga. All letters should be addressed to the Atlanta office, 402 Trust Co. of Ga. Bldg.
Southern Architect and Contractor, Atlanta, Ga., and Southern Building News, Memphis, Tenn., and Southern Building Record, Nashville, Tenn., Concrete Age, Atlanta, Ga., Consolidated.

1924 SCHEDULE FOR ADVERTISERS

For Jan. and Feb. issue copy and all cuts must reach us by ..... Dec. 20th
For March ..... Jan. 20th
For April ..... Feb. 20th
For May ..... Mar. 20th
For June ..... April 20th
For July ..... May 20th
For August ..... June 20th
For Sept. ..... July 20th
For Oct. ..... Aug. 20th
For Nov. ..... Sept. 20th
For Dec. ..... Oct. 20th



STORE FRONTS

KAWNEER "B" CONSTRUCTION

A modern frame for modern merchandise ... Simple in construction ... Surfaces unblemished by rivets or screws ... Strong and durable ... Easily installed, presenting a unified appearance in the finished front ... Made of rolled or extruded Bronze or Aluminum Alloy. Write for Full Size Details and New Catalog "N."

THE Kawneer COMPANY
NILES, MICHIGAN
FACTORIES
NILES • CHICAGO HEIGHTS
CHICAGO • BERKELEY

PATENTS AND COPYRIGHT APPLIED FOR

The New FA Fusebox for Residences

The newest member of the great FA line of Panelboards is this "just out" FA Fusebox. It is designed for use in small houses and wherever else a compact unit is needed.

It is built of molded sections, safety type, with or without branch circuit switches. An inch and a half gutter gives ample wiring space. A white enameled door and a front, cut from one piece of sheet metal, makes it "the sign of a better job" wherever installed.

This FA Fusebox is priced in accordance with its economy in manufacture and will be welcome to all interested in small home building.

Send for complete details and estimates

Frank Adam
ELECTRIC COMPANY
ST. LOUIS

DISTRICT OFFICES

Atlanta, Ga., Baltimore, Md., Boston, Mass., Buffalo, N. Y., Chicago, Ill., Cincinnati, Ohio, Dallas, Texas, Denver, Colo., Detroit, Mich., Jacksonville, Fla., Kansas City, Mo., Los Angeles, Calif., Memphis, Tenn., Minneapolis, Minn., New Orleans, La., New York, N. Y., Omaha, Nebr., Philadelphia, Pa., Pittsburgh, Pa., San Francisco, Calif., Seattle, Wash., Tulsa, Okla., Montreal, Que., Toronto, Ont., Vancouver, B. C., Walkerville, Ont., Winnipeg, Man.

Entered as second-class matter January 15, 1928, at the Post Office, Atlanta, Ga., under Act of March 3, 1879
General Office, 434-436 Trust Company of Georgia Building, Atlanta, Ga.

Illustrated monthly architectural trade publication aimed at audience of industry professionals

Published between 1889-1932

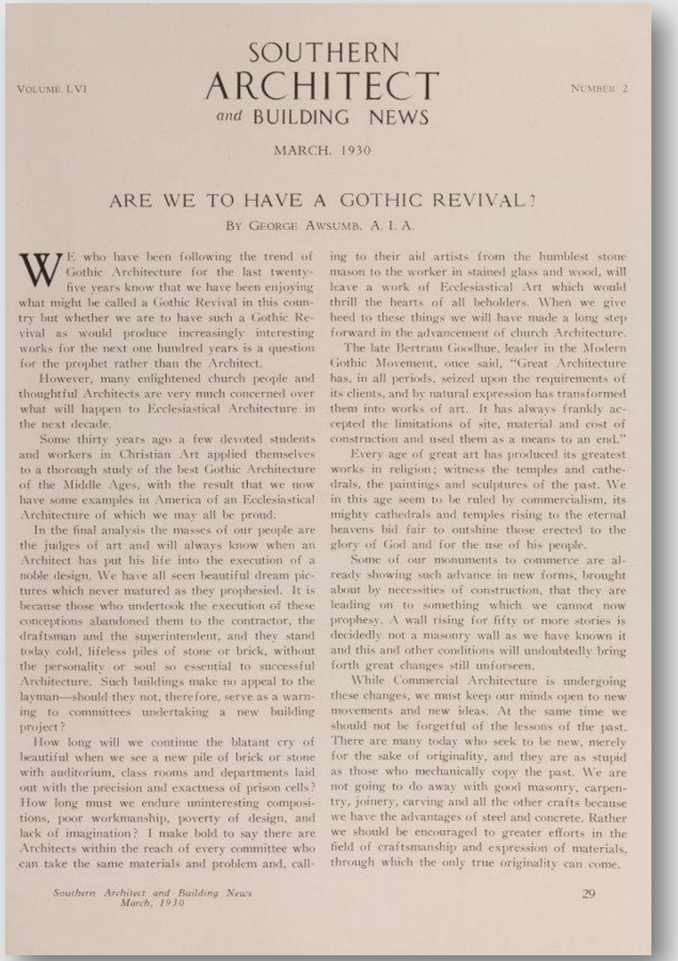
Held in Special Collections, John S. Chase Architecture and Planning Library

254 distinct issues = 31,250 pages

Pictured: Contents Page, Dual-Column Advertisement

Challenge to determine relationships between elements, and to understand dated formatting of schedule for advertisers

Why this collection?



VOLUME LVI — SOUTHERN ARCHITECT and BUILDING NEWS — NUMBER 2

MARCH, 1930

ARE WE TO HAVE A GOTHIC REVIVAL?

By George Awsumb, A. I. A.

WE who have been following the trend of Gothic Architecture for the last twenty-five years know that we have been enjoying what might be called a Gothic Revival in this country but whether we are to have such a Gothic Revival as would produce increasingly interesting works for the next one hundred years is a question for the prophet rather than the Architect.

However, many enlightened church people and thoughtful Architects are very much concerned over what will happen to Ecclesiastical Architecture in the next decade.

Some thirty years ago a few devoted students and workers in Christian Art applied themselves to a thorough study of the best Gothic Architecture of the Middle Ages, with the result that we now have some examples in America of an Ecclesiastical Architecture of which we may all be proud.

In the final analysis the masses of our people are the judges of art and will always know when an Architect has put his life into the execution of a noble design. We have all seen beautiful dream pictures which never matured as they prophesied. It is because those who undertook the execution of these conceptions abandoned them to the contractor, the draftsman and the superintendent, and they stand today cold, lifeless piles of stone or brick, without the personality or soul so essential to successful Architecture. Such buildings make no appeal to the layman—should they not, therefore, serve as a warning to committees undertaking a new building project?

How long will we continue the blatant cry of beautiful when we see a new pile of brick or stone with auditorium, class rooms and departments laid out with the precision and exactness of prison cells? How long must we endure uninteresting compositions, poor workmanship, poverty of design, and lack of imagination? I make bold to say there are Architects within the reach of every committee who can take the same materials and problem and, calling to their aid artists from the humblest stone mason to the worker in stained glass and wood, will leave a work of Ecclesiastical Art which would thrill the hearts of all beholders. When we give heed to these things we will have made a long step forward in the advancement of church Architecture.

The late Bertram Goodhue, leader in the Modern Gothic Movement, once said, "Great Architecture has, in all periods, seized upon the requirements of its clients, and by natural expression has transformed them into works of art. It has always frankly accepted the limitations of site, material and cost of construction and used them as a means to an end."

Every age of great art has produced its greatest works in religion; witness the temples and cathedrals, the paintings and sculptures of the past. We in this age seem to be ruled by commercialism, its mighty cathedrals and temples rising to the eternal heavens bid fair to outshine those erected to the glory of God and for the use of his people.

Some of our monuments to commerce are already showing such advance in new forms, brought about by necessities of construction, that they are leading on to something which we cannot now prophesy. A wall rising for fifty or more stories is decidedly not a masonry wall as we have known it and this and other conditions will undoubtedly bring forth great changes still unforseen.

While Commercial Architecture is undergoing these changes, we must keep our minds open to new movements and new ideas. At the same time we should not be forgetful of the lessons of the past. There are many today who seek to be new, merely for the sake of originality, and they are as stupid as those who mechanically copy the past. We are not going to do away with good masonry, carpentry, joinery, carving and all the other crafts because we have the advantages of steel and concrete. Rather we should be encouraged to greater efforts in the field of craftsmanship and expression of materials, through which the only true originality can come.

NORTHINGTON RESIDENCE, BIRMINGHAM, ALA.
WARREN, KNIGHT & DAVIS, ARCHITECTS.

Pictured: Residential Architecture Photograph, Text-Heavy Editorial

Challenge to assign meaning when there is very little text, and to parse meaning from so much dense text

Our collection is fully digitized, but not as discoverable as we would like

Ongoing interest and a project with the School of Architecture to index collection and identify entities and themes

OCR created for that project initially. It contains errors that are a result of unique formatting and the limitations of OCR technologies - AI has the potential to improve the quality of OCR texts

Overview: Project Structure

RESEARCH DESIGN	• Implementation focused • Web UI & API • Widely Available Models • Varied Input Methods
TASKS	• Tasks are useful, common, time consuming • Subject heading assignment, content summarization, entity extraction, OCR text transcription, content warning generation, visual description (image workflow)
EVALUATION PRACTICES	• Team members had a variety of perspectives • Blind evaluations to decrease bias • Analyzed content scores, tester patterns, time, cost, and other factors

02

Testing & Assessment Methodology

Prompt Design

Adapted from the PROMPT design framework (Hartman-Caverly, 2024)

Iterative improvement by testing different strategies

Specialized prompts for types of content (covers and image pages)

Who (role)
What (context)
Why (goals)
How (instructions)

Persona: Archivist at the University of Texas at Austin with expertise in…
Requirements: Do not alter content, do not alter names, do fix punctuation, etc.
Organization: "Please create the following metadata fields:"
Medium: Describe format of output: JSON
Purpose: Metadata for architecture students and scholars, prioritize architectural styles, etc.
Tone: Archival standard (ex: OCR be written exactly as on the page without changing language)

Assessment Criteria

Completeness

Do the description and metadata terms fully describe the content?

Accuracy

How clean is the OCR? Do terms & headings exist? Are the ToC and description accurate?

Usefulness

Would the metadata help the searcher find the material? Is it related to the subject?

Testing Methodology

API

Python scripts developed in fall 2024 to automate processing

LLM responses parsed and automatically written to spreadsheets

Text: December 9, 2024
Image: December 11-12, 2024

Claude Haiku, Claude 3 Sonnet, Claude Sonnet 3.5, GPT-4o-mini, GPT-4o, Gemini 2.0 Flash

Web

Tested using free accounts, exception of Copilot subscription

Prompts divided due to account limitations.
Results manually compiled.

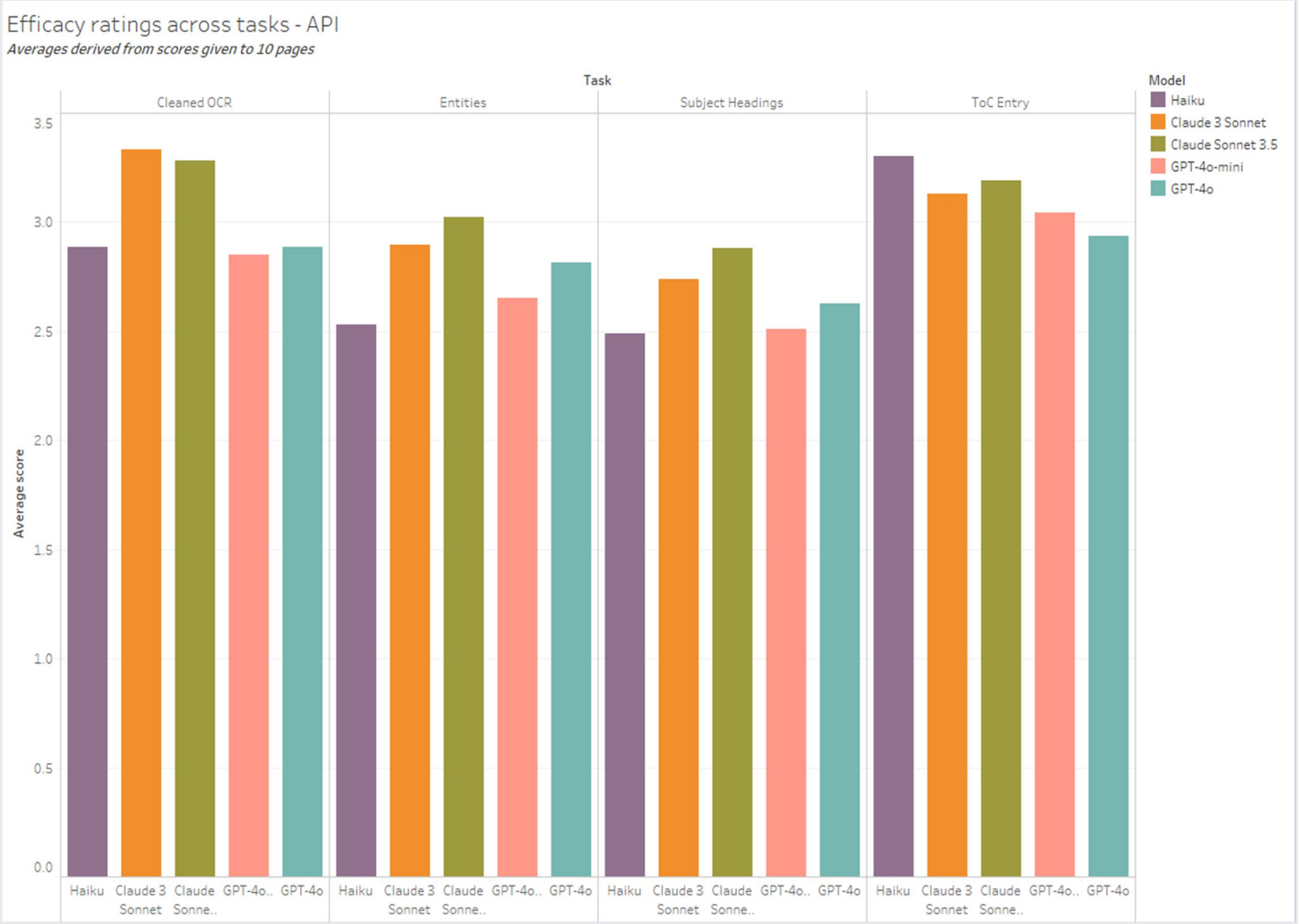
Text: October 2024
Image: February 2025

Claude Sonnet, CoPilot, GPT-4o-mini, Gemini, Llama

03

Results & Analysis

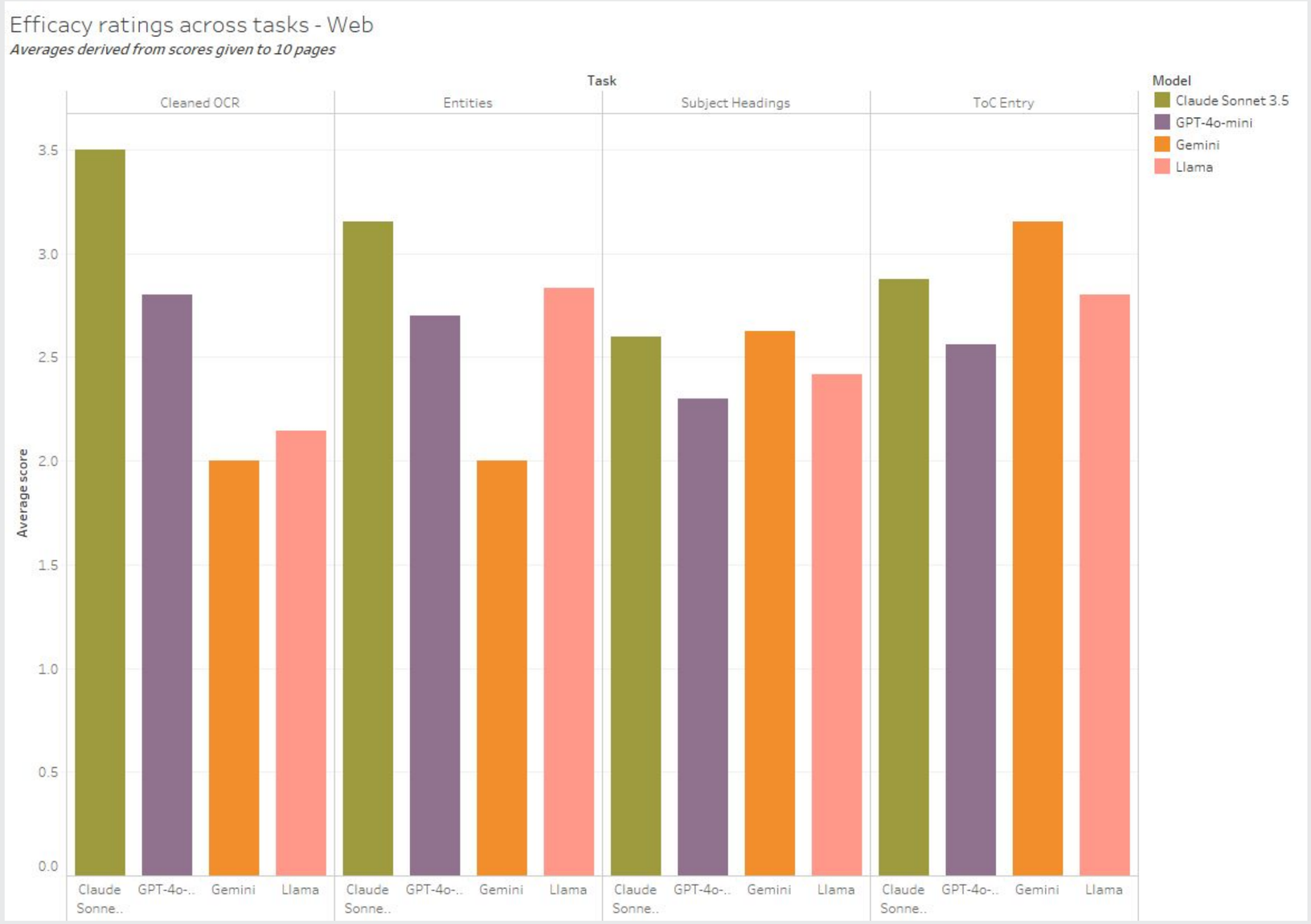
API - Text Workflow



API : Haiku, Claude 3 Sonnet, Claude Sonnet 3.5, GPT-4o-mini, GPT-4o

- Claude Sonnet 3.5 scored highest across API and Web workflows
- API workflow had more consistent results than the Web workflow
- Best in Task:
  - Cleaned OCR: Claude 3 Sonnet
  - Entity Extraction: Claude Sonnet 3.5
  - Subject Headings: Claude Sonnet 3.5
  - Summary (ToC Entry): Haiku (a smaller model!)
- Subject Heading Generation was the lowest scoring task across API and Web workflows
- Named entity scores were generally low as well, but testers still found the results useful

Web - Text Workflow



Web: Chat GPT 4 mini, Gemini 3, Claude, Llama

- Less consistent results overall than working with API
- Best in Task:
  - Cleaned OCR: Claude Sonnet 3.5
  - Entity Extraction: Claude Sonnet 3.5
  - Subject Headings: Gemini
  - Summary (ToC Entry): Gemini
- Performed well overall in extracting named entities
- Gemini's performance was particularly unreliable – it got the lowest scores in entity extraction and OCR editing

Image Workflow Results

API

- Scored higher than text workflow for OCR creation and entity extraction
- Summarization similar to text workflow
- Subject heading creation very slightly lower
- Overall, the highest scores of all the workflows

Models: Claude Haiku, Claude 3 Sonnet, Claude Sonnet 3.5, GPT-4o, GPT-4o-mini, Gemini 2.0 Experimental

WEB

- Text and image workflows on the web had problems with reliability
- Scored lower than text workflow across all tasks
- More frequently returned shortened versions of OCR texts
  - Possibly due to context limitations in free versions of LLMs

Models: Claude Sonnet, CoPilot, GPT-4o-mini, Gemini, Llama

Sentiment Analysis

Methodology

- Annotations compiled in R Studio and analyzed using sentimentr package

Blank fields

- Scored as neutral; analysis performed with and without blank fields

Annotation fatigue

- Models scored later in process included more blank fields – testers likely grew tired of noting reasoning

Individual rating Styles

- Testers scored LLM's differently on an absolute scale, with some giving higher or lower scores across the board.

Consistent relative patterns

- Relative sentiment expressed for different models and tasks was consistent across testers

04

Other Implementation Considerations

Variation in LLM Responses to Offensive Content

API

- Text Inputs:
  - All models processed page
  - Claude Sonnet 3.5 replaced offensive words with neutral words
- Image Inputs:
  - Gemini and GPT-4o-mini refused to process page
  - Claude 3 Sonnet and Haiku used offensive terms in unnecessary metadata fields
- GPT-4o was the only model that handled content well with both text and image inputs – transcribed, didn't repeat unnecessarily, and noted that there was potentially harmful language

WEB

- Image-based Web:
  - Claude Haiku : "The article contains academic discussion and critique of sculptural art from a historical perspective. No overtly sensitive content was identified."
  - ChatGPT: "The historical article includes period-specific language and perspectives that may reflect biases or assumptions of The era."

Practical Considerations

Time

- Larger models take longer, input doesn't affect time as much
  - Per 1,000 pages:
    - GPT-4o: 5.5 hours
    - GPT-4o-mini: 2 hours
- Typically, batch processing guarantees a processing time of 24 hours

Input Format

- API: image inputs consistently scored highest
- Cost differences vary by model
- Per 1,000 pages:
  - Haiku is $.48 more for image inputs
  - Claude Sonnet 3.5 is $2.98 more for image inputs

Cost

- Larger models and image processing cost more.
  - Per 1,000 pages:
    - Claude Sonnet 3.5 image: $14
    - GPT-4o-mini image: $1.15; text: $0.52
    - Haiku image: $1.34
- Batch processing can cut costs in half

05

Recommendations

Pilot Project Design

1

Assess Collection

Is this collection is a good fit for an AI workflow?
Think about ethics, rights, format, content, etc.
Decide on metadata needs and how AI-Generated metadata will fit into your current systems and processes.

2

Define Success Criteria

What makes metadata 'good' for this collection and the tasks we have set?
What threshold do you need to pass in order for this work to be valuable?

3

Gather Team

Involve people in the assessment who have different backgrounds and perspectives and a stake in the content you're producing

4

Test Models & Methods

Using LLMs that are available to you (free/institutional web, API), test with text and image inputs. Try different prompting approaches. Track time (if applicable) and costs.

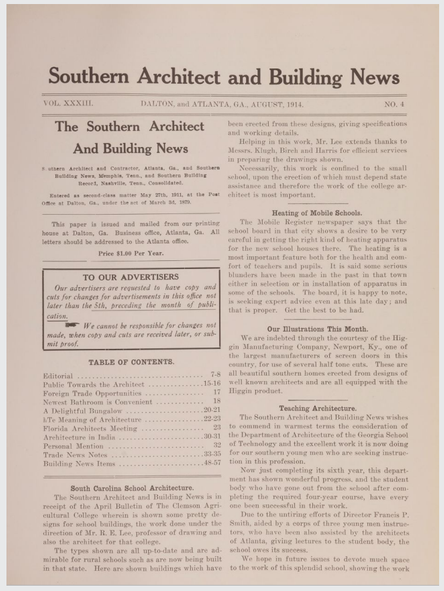
5

Analyze findings & Plan Approach

Choose the best model for your collection and tasks, decide if you need to incorporate grounding steps (other APIs), determine best practices for quality control, and document recommended workflow.

Next Steps: Southern Architect & Building News Collection

- Page-level metadata is at this stage too detailed – we're just not sure where to put it all!
- Plans for the metadata: replace issue-level OCR for use in out Collections portal and From the Page
- Tasks: OCR text transcription, summarization, keyword creation, entity extraction
- Model: GPT-4o (have access to an institutional OpenAI account)
- Tasks incorporating other APIs: subject heading search using the keywords, entity validation/search
- Will test using an intentionally chosen subsection of each journal issue
- Pages that we will use: cover, table of contents, table of advertisers, and pages that contain architect and building information, such as "Building Notes," "Building and Construction News," and "Building and Construction Department"



Southern Architect and Building News

VOL. XXXIII. DALTON, and ATLANTA, GA., AUGUST, 1914. NO. 4

The Southern Architect And Building News

TABLE OF CONTENTS.

South Carolina School Architecture.

Heating of Mobile Schools.

Our Illustrations This Month.

Teaching Architecture.

THE SOUTHERN ARCHITECT AND BUILDERS' HARDWARE JOURNAL

Vol. IV. No. 12. ATLANTA, GA., OCTOBER, 1893. $2 A YEAR; SINGLE COPIES 20C

HIGH SERVICE ELEVATORS.

Hydraulic Passenger Elevators, Hydraulic Freight Elevators. Worm-Geared Elevators,
Hydraulic Variable Lift Elevators. Hand-Power Carriage Elevators. Dumb-Waiters.
Hydraulic Plunger Elevators, Hand-Power Freight Elevators.

THE JAMES L. HAVEN COMPANY, CINCINNATI, OHIO.

THE OTIS Elevator

THOUSANDS IN USE.

OTIS BROTHERS & CO., 38 Park Row, New York. H. E. W. PALMER, Gen'l Agt. Atlanta Office, Equitable Building.

TENNESSEE MARBLE!

INTERIOR FINISH. BUILDING MARBLE.

Atlanta Elevator Co. —MAKERS— Passenger and Freight ELEVATORS.

OFFICE AND WORKS, Means Street, W. & A. R. ATLANTA, GA.

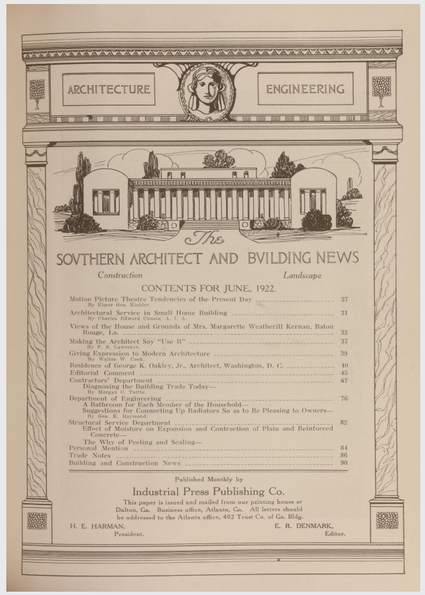
D. C. BACON, President. A. C. ARGUIMBAU, Gen. Sec.

Atlanta Lumber Co.

GEORGIA YELLOW PINE LUMBER, SHINGLES, LATHS.

Special Sizes Cut for Buildings, Bridges, Trestles and Cars.

ATLANTA, GA.



ARCHITECTURE ENGINEERING

The SOUTHERN ARCHITECT AND BUILDING NEWS

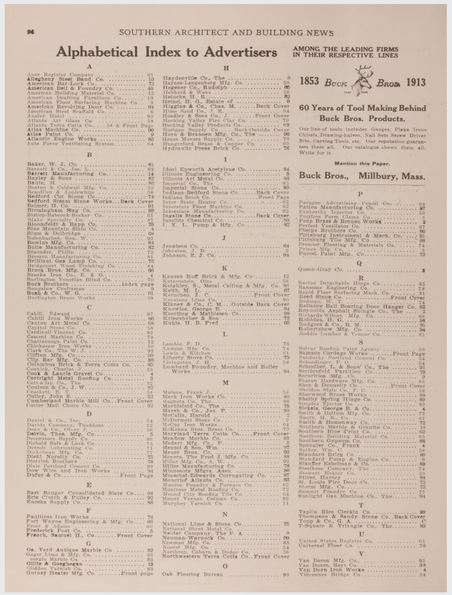
Construction Landscape

CONTENTS FOR JUNE, 1922.

Published Monthly by Industrial Press Publishing Co.

This paper is issued and mailed from our printing house at Dalton, Ga. Business office, Atlanta, Ga. All letters should be addressed to the Atlanta office, 402 Trust Co. of Ga. Bldg.

H. E. HARMAN, President. E. R. DENMARK, Editor.



SOUTHERN ARCHITECT AND BUILDING NEWS

Alphabetical Index to Advertisers

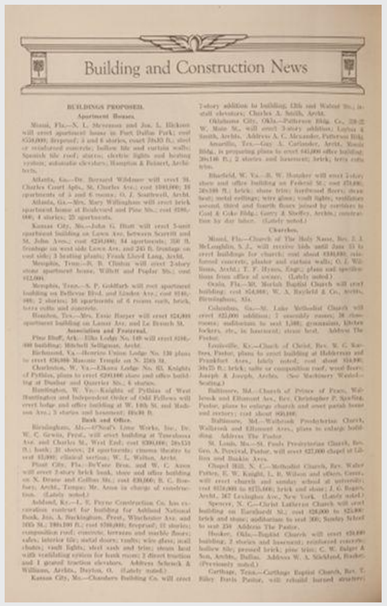
AMONG THE LEADING FIRMS IN THEIR RESPECTIVE LINES

1853 Buck Bros. 1913

60 Years of Tool Making Behind Buck Bros. Products.

Mention this Paper.

Buck Bros., Millbury, Mass.



Building and Construction News

BUILDINGS PROPOSED.

Apartment Houses.

Churches.

Association and Fraternal.

Bank and Office.

Recommendations

Budget-Conscious Implementation

- Use free or paid versions of chatbots on web
- API access less than you might think – worth researching
- Use smaller models like GPT-4o-mini or Haiku and take advantage of batch processing
- No matter what:
  - Iteratively improve prompts
  - Incorporate grounding and/or validation
  - Choose AI models and tasks based on results obtained
  - Design for your collection

Quality Focused Implementation

- Winner: Claude Sonnet 3.5
- If your collection contains problematic language, consider GPT-4o
- Still take advantage of batch processing
- Build workflows that use the best models for each task – for API for example, it might be:
  - Original OCR texts to Claude Sonnet 3.5 for transcription
  - OCR text to Haiku for summarization and keyword suggestion,
  - Keywords can be used as search terms for a FAST API

Hybrid Approach

- Be sure that batch processing is available
- Start your workflow design by thinking about what you will do with the metadata you generate
- Choose models that are more expensive only when it increases quality. Via API, OCR and entity extraction were better with image inputs, but it made no difference for summarization.
- Test, iterate, test again, run, and establish quality control practices

Thank You!

Get in touch with questions, comments, ideas, or just to chat about AI and metadata!

- Hannah Moutran: hlm2454@my.utexas.edu
- Karina Sanchez: karinasanchez@austin.utexas.edu
- Katie Pierce Meyer: katiepiercemeyer@austin.utexas.edu
- Devon Murphy: devon.murphy@austin.utexas.edu
- Willem Borkgren: willem.borkgren@austin.utexas.edu
- Josh Conrad: j.conrad@austin.utexas.edu

Special thanks to Aaron Choate, Director of Research & Strategy at UT Libraries

Presentation Template: Slides Carnival, photos/images Pexels & Pixabay