BENEFITS INVESTIGATOR

Center for Enterprise Dissemination (CED) | U.S. Census Bureau

Brandon Pardi, Data Science Fellow | *University of California, Davis, Computer Science M.S.*

Keywords:

Language Models, Pipeline, API, SQL Databases, Python

Summary:

Researchers submit proposals to gain access to Census data, indicating which out of 13 possible benefits they will provide. Brandon Pardi worked with the FSRDC to develop a project with a clean, expandable pipeline for retrieving, parsing, and cleaning these documents, and analyzing via the GPT API. In a study, 40 legacy publications and 20 working papers were quickly processed, with AI identifying more diverse and applicable benefits than human researchers. The findings suggest that while GPT-4-turbo and other large language models (LLMs) are effective, they can be fine-tuned for specific tasks. Additionally, free LLMs like Meta's Llama 3 can securely process restricted data with the right hardware.

BENEFITS INVESTIGATOR

How are Researchers Using Census Data Benefitting the Bureau?

Center for Enterprise Dissemination (CED)

U.S. Census Bureau

Supervisors: Jay Kriebel, David Beede, Ashley Landreth

coding it forward >

United States™ Census Bureau

Brandon Pardi

University of California, Davis

Computer Science M.S.

WHY DOES THE FSRDC EXIST?

- Federal Statistical Research Data Centers program exists within the CED.
- Facilitates the access of restricted microdata from various statistical governmental agencies to screened researchers.
- Want to keep track of research results to understand the data's impact.

PROJECT OVERVIEW

- Researchers submit proposals to gain access to Census data.
- 13 possible benefits that researchers indicate they can provide.
    - Data shows typically 1-3 benefits are indicated per project.
- Are these benefits actually being fulfilled? Are more being fulfilled than initially indicated?

cif>

WHAT CONSTITUTES A BENEFIT?

Example Benefit:

Criterion 5: The project benefits the Census Bureau by helping to understand or improve the quality of data the Census Bureau collects or acquires.

Explanation: The Census Bureau needs to understand and continually assess the quality of all the data it collects, and to seek ways to improve them. Understanding the limitations of and improving the quality of these input data can greatly expand their utility and the quality of the analyses they inform.

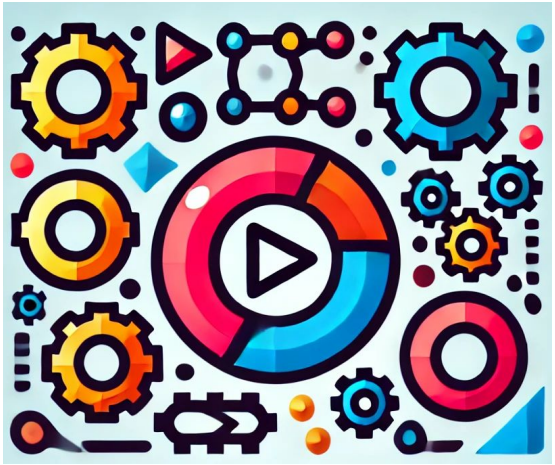
cif>

THE BENEFITS PIPELINE

How we analyze a research paper's benefits

cif>

WHAT I STARTED WITH



Confusing Codebase

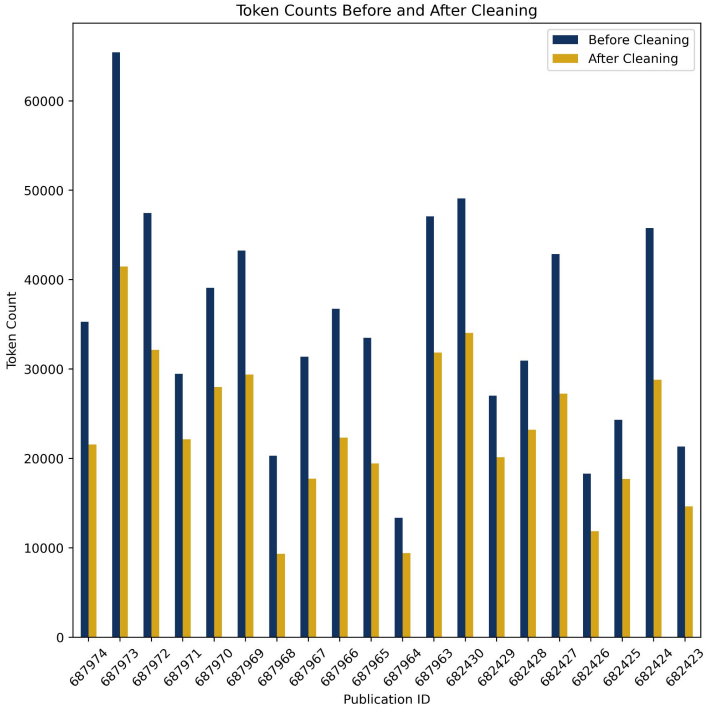
SQL Database

A Dream

cif>

WHAT I ENDED WITH

- Clean, easily expandable pipeline for retrieving, parsing, and cleaning documents, to send to GPT API and store the results
- Analyzed the benefits in 40 legacy publications and 20 working papers
- Papers range from 15 - 120 pages.
- All were processed in less than 20 minutes.



Results

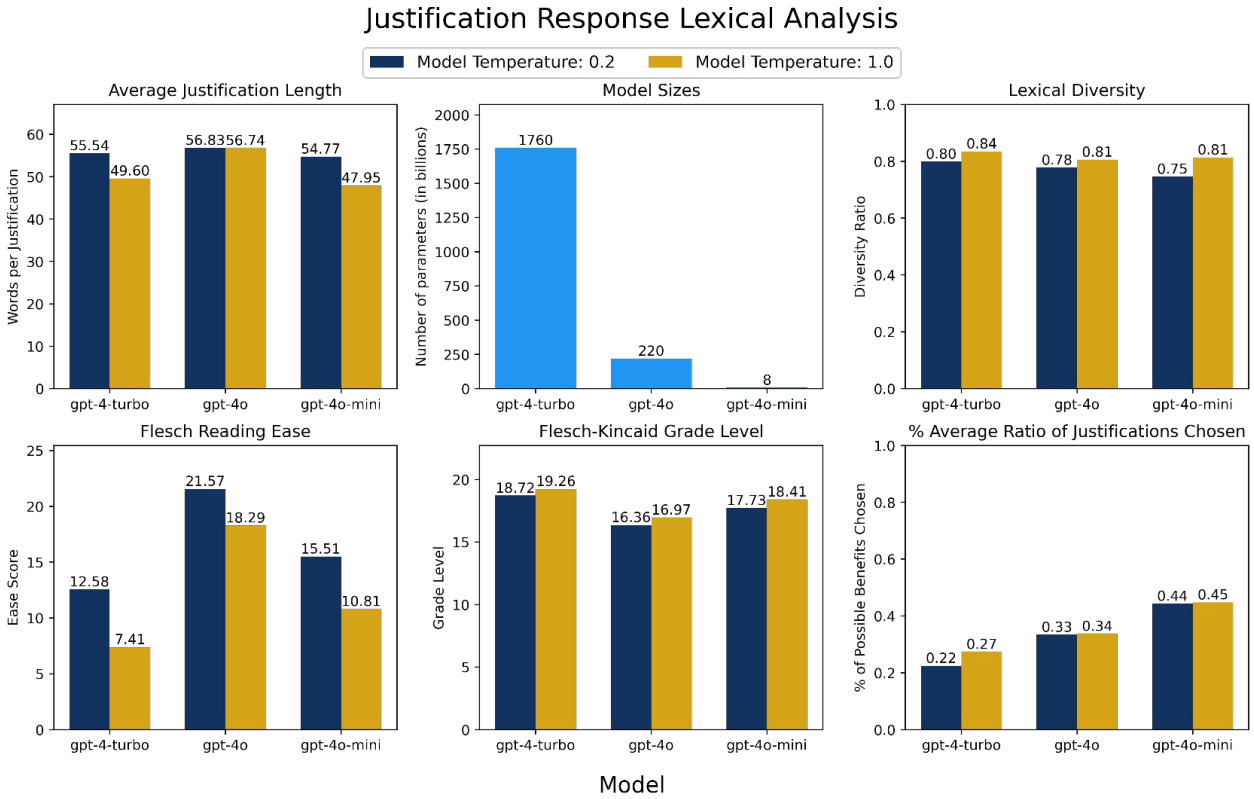
How well did we do?

cif>

HUMAN VS AI

- 20 working papers processed by GPT-4-turbo.
- Researchers generally chose the same 3 benefits.
    - All of them chose at least benefit 11.
- GPT was more diversified and generally chose more applicable benefits than researchers did.
- Decent amount of overlap

MODEL COMPARISONS



A CALL TO ACTION

How can we improve this?

cif>

IN HOUSE LANGUAGE MODELS

- LLM's are smart but are trained on generic data.
  - Can be fine tuned.
- Free LLM's exist comparable to GPT.
  - Meta's Llama 3.
  - No need to develop a LLM from scratch.
- Can process restricted data with NO security risks.
- All that is needed are some GPUs

cif>

THANK YOU

Special thanks to Jay Kriebel, Ashley Landreth, David Beede, and the CIF team!

cif>