BUILDING TRUST IN AI FOR QUALITATIVE DATA ANALYSIS

U.S. Census Bureau | Economic Statistical Methods Division

Winston Li | University of California, Los Angeles; *Statistics and Data Science & Political Science*

Keywords: Artificial intelligence, large language models, retrieval augmented generation (RAG), recursive abstractive processing for tree-organized retrieval (RAPTOR), natural language processing

Technologies: Python, TensorFlow, Hugging Face, LlamaIndex, Milvus, Pandas, CUDA

Summary: I developed an offline, private chat product to analyze Census reports while addressing challenges with contextual depth, algorithmic bias, and model hallucinations. To enhance response accuracy, I constructed an ensemble pipeline that effectively synthesizes findings across documents using two models. I also implemented an advanced recursive AI retrieval system to improve contextual understanding versus industry solutions by preprocessing, chunking, and ranking documents. Finally, I created an automated, purpose-built testing framework to benchmark and compare model performance in terms of accuracy, reliability, and relevance.

The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data (Project No. P-7530157, Disclosure Review Board (DRB) approval number: CBDRB-FY23-ESMD010-013).

Building Trust In AI For Qualitative Data Analysis

By: Winston Li
Economic Statistical Methods Division

Supervisor: Rebecca Keegan, Survey Statistician

The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data (Project No. P-7530157, Disclosure Review Board (DRB) approval number: CBDRB-FY23-ESMD010-013).

OBJECTIVES

- Create an offline, open-source generative AI chatbot to:
  - Answer questions about the content of documents
  - Synthesize findings between documents to draw larger conclusions
  - Generate potential questions
  - Determine relevant codes from documents
- **Need to improve reliability, reduce hallucinations versus current models**
- AI can potentially be more **accurate**, **effective**, and **objective** in identifying **thematic connections** within and between qualitative reports and documents

Example questions:

1. What are some common difficulties users had when interacting with Census surveys?
2. Within F&R reports, what are some of the common themes found when examining questions? What types of questions were found to be poorly designed for users and why?
3. Within these reports, what types of tasks and questions are respondents finding burdensome? Additionally, please report which types of tasks and questions are of low burden to the respondent.

UNDERSTANDING MODELS

Chatbot Context

"Short-term memory"

Limited in input length & expensive

Understanding of what it is provided

Retrieval-Augmented Generation (RAG)

"Open-note exam"

Ability to search many documents

Limited understanding of full context or meaning

Recursive Abstractive Processing for Tree-Organized Retrieval (RAPTOR)

Pre-processing to understand full context and relationships

Ability to search many documents

Slower, more resource-intensive

PIPELINE

1

Load documents

Documents are provided to the AI model to pre-process and index.

2

Ask question

User asks a question to the chatbot.

3

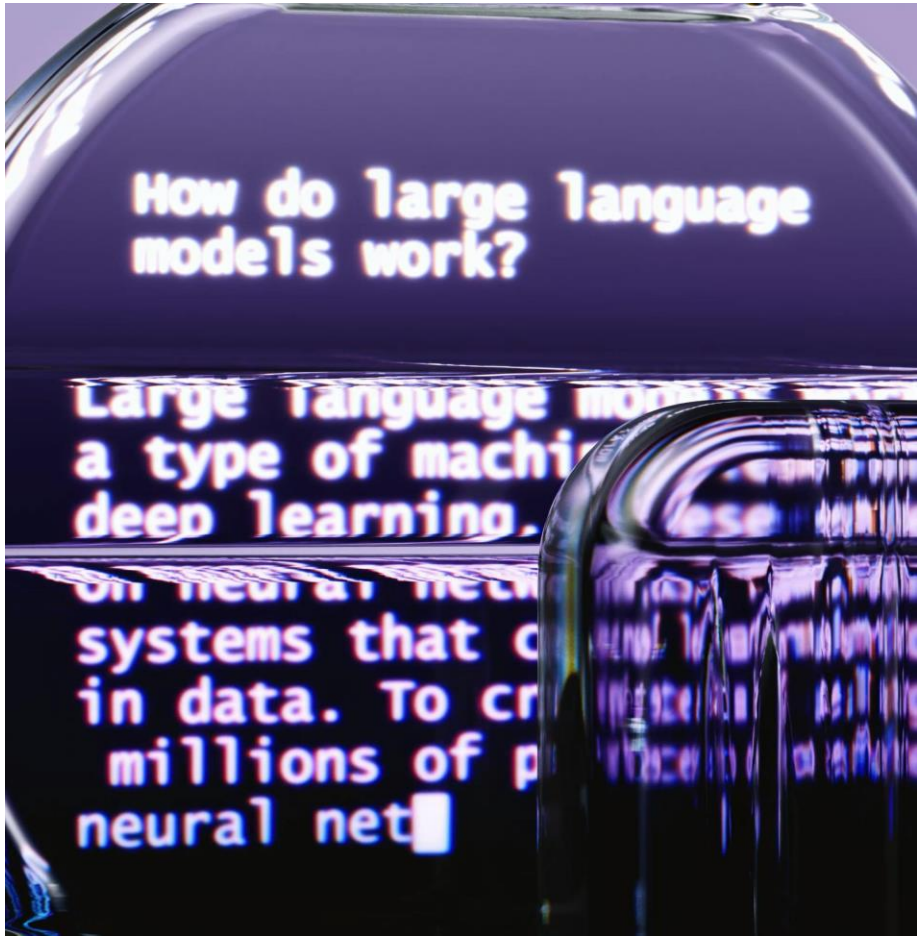
Generate answer

The main AI model answers the user's question.

4

Fact check

The secondary AI model fact checks the main model's answers with citations.



TESTING

United States Census Bureau

Choosing an AI model that performs consistently well across fact check models.

Main Model ➡️ Fact Check Model ⬇️	OpenAI GPT-4 (closed-source)	Google Gemini 1.5 Pro (closed-source)	Microsoft Phi 3 (open-source)	**Mistral v0.3 (open-source)**	Meta Llama 3 (open-source)	Berkeley-NEST Starling (open-source)
♊ Gemini 1.5 Pro	🟢 0.86	🟢 **0.87**	🟡 0.83	🟡 0.83	🟡 0.81	🔴 0.78
Φ Phi 3	🟢 0.80	🟡 0.78	🟡 0.72	🟢 **0.9**	🟡 0.74	🔴 0.66
🍃 Mistral 0.3	🟢 0.86	🟢 **0.91**	🔴 0.67	🟡 0.79	🟡 0.7	🟢 0.86
🦙 Llama 3	🟡 0.64	🔴 0.51	🟡 0.62	🟢 **0.8**	🟡 0.73	🟢 0.76
💫 Starling LM	🟡 0.82	🟢 0.85	🟡 0.67	🟢 **0.89**	🔴 0.53	🟡 0.73

Example ensemble (main + fact-check): Mistral v0.3 + Microsoft Phi 3

- GPT-4 was not used as a fact-checking model due to issues with inconsistent/illogical fact-checks in testing
- Meta Llama 3.1 was initially tested but not used due to infinite generation issues

INTERFACE

Home

- Home
- RAG Chat
- Read Me!

RAPTOR (experimental)

- Step 1: RAPTOR Database
- Step 2: RAPTOR Chat

Model Testing (advanced)

- Step 1: Preparation
- Step 2: Generation
- Step 3: Scoring

RAG: No pre-processing required

RAPTOR: Requires pre-processing, better contextual understanding

Model Testing and Comparison: Prepare your questions, generate initial answers, fact-check and score

Hello!

This app is a proof of concept for how offline, open-source large language models can be used to analyze complex qualitative documents and draw thematic connections.

If the top right corner says ' Running', do not click any items on the page or switch pages as it may interrupt the current task.

Please feel free to watch the two videos in the folder (**Part 1 - User Experience.mp4** and **Part 2 - Hardware, Code.mp4**) for more information and context!

Scripts

RAG Chat: This is a standard retrieval augmented generation (RAG) chatbot with automatic fact-checking capabilities. It is ideal if you are looking to quickly analyze documents and find the 'needle in the haystack', but it is weaker at understanding the broader contexts of documents.

Recursive Abstractive Processing for Tree-Organized Retrieval (RAPTOR)

Step 1: RAPTOR Database: If you wish to use recursive processing (RAPTOR), which is better at understanding concepts between large numbers of documents, you must first create a RAPTOR database on the documents you want to analyze. This does take a while depending on how much text you are analyzing.

CHAT

United States Census Bureau

- Home
- Standard RAG
- Pre-Process RAPTOR
- Advanced RAPTOR
- Read Me!

About

AI Chatbot with real-time fact-checking

File Management

Default files on the server are located under read_reports. If you wish to upload files, add them under

Choose a file

Drag and drop file here

Limit 200MB per file • TXT, PDF, DOCX

Browse files

(1) Load documents

(2) ❓ Ask question

Within these reports, what are some of the common themes found when examining questions? What types of questions were found to be poorly designed for users and why?

(3) AI answers

(4) Fact-check each segment

1. The reports indicate that some questions in the surveys may be perceived as overwhelming due to the amount of content and the number of questions asked. Participants suggested ways to lessen this feeling, such as scaling back the content, breaking up the spreadsheet by topic, and using separate spreadsheets, tabs, or pages to organize topics.

Fact-check: True. Page 11 of the document titled "Findings and Recommendations from Usability Vignette Testing for the Annual Integrated Economic Surveys" states: "Many suggested breaking up the spreadsheet in a variety of ways, mostly by topic. They offered ideas such as separate spreadsheets, tabs or pages to organize topics..." (Document: Findings and Recommendations from Usability Vignette Testing for the Annual Integrated Economic Surveys, Page 11).

2. The NAPCS section in the AIES spreadsheet was found to be visually overwhelming and a lot to look at for some participants.

3. Some topics, such as reshoring, labor, entrepreneurship, SOGI, and disability, were found to be poorly designed for users and may benefit from a different approach to gathering information, such as open-ended interviews, focus groups, usability testing, or respondent debriefing interviews, to ensure the research question matches the method. This is to ensure that the questions are appropriate and relevant for the companies being surveyed and to assess the perceived sensitivity, reaction, impression of the questions and response options, level of willingness to report the

What's on your mind?

CONCLUSION

- AI models have improved vastly since last year
- **Strengths:** identifying patterns, finding information
- **Weaknesses:** full understanding, speed versus accuracy
- **Recommendations:** fact-check answers, manually check results

THANK YOU!