Efficacy of AI RAG Tools for Complex Information Extraction and Data Annotation Tasks on Bank Documents

Nicholas Botti (Presenter), Flora Haberkorn, Shaun Khan

Board of Governors of the Federal Reserve System

Introduction

- There is tremendous interest in utilizing Large Language Models (LLMs) and related techniques such as retrieval augmented generation (RAG) to extract information from financial documents and other types of unstructured data
- Some general benchmarks exist for this type of task, such as RAGAS (Es et. al. 2023) or needle in a haystack (Kamradt 2023)
- We apply these techniques to a **challenging, real-world information extraction task** on bank documents that took multiple human annotators months to complete (Beltran et. al. 2023)

Structured Data Annotation Task

- Using dataset created by (Beltran et. al. 2023)
- Large global banks report information about their actions, policies, commitments, etc. relating to various topics in their annual reports and other public disclosure documents
- However, because there is no structured dataset of this information, it is difficult to look at industry-wide trends, etc.
- Task objective: Have annotators review these unstructured text data sources, and qualitatively code variables of interest to create a high-quality dataset for researchers

Note: Information obtained during this project, or during the original research paper, is not used for supervisory purposes by the Federal Reserve Board

A Unique Challenge for AI

- ~1000-2000 pages worth of documentation to review per bank
- Complex multi-part questions require combining/cross-referencing multiple sections of text
- Heavy usage of finance domain-specific terminology
- ~150 questions to answer per bank.
- Not all questions have answers in the provided documents.
- Some questions require subjective interpretation.
- **Research goal: understand how effectively AI tools can assist human annotators on this type of complex task**

Experiment Methodology

- Select a representative subset of 30 questions from the original task, and select a subset of nine European banks
- Randomly assign half of the questions (15) to the control group, and half (15) to the treatment group for each bank
- Use a pre-selected group of documents for each bank
- Annotators given general background on the task, but no extensive training. Annotators recorded time to answer each question
- Control group: annotators answer the questions using only the provided documents
- Treatment group: annotators given access to an AI tool, pre-populated with the set of documents for their bank. They answer based ONLY on the information provided by the tool, and are instructed not to look at the documents themselves

AI Tool Usage Procedure

- Annotators were given access to an AI document chat interface
- Wrote their own prompts based on the question
- Used their judgement to qualitatively code and record answer according to provided criteria
- Encouraged to send only one prompt per question
  - Could ask the tool up to one additional follow-up question if necessary

AI Tool Architecture

- Used a basic retrieval augmented generation (RAG) approach
  - Embedding Model: Amazon Titan Embed Text v2 (1024 dimension). 2000-character passages with 100-character overlap.
  - Top 20 most relevant passages sent as context along with prompt to LLM
  - LLM: Claude 3.5 Sonnet (20240620 version) via AWS Bedrock
- Note: This is not a state-of-the-art RAG approach
  - Objective of this research was to establish the feasibility of AI tools on this type of task using tools commonly available to researchers
  - Models and architecture consistent for all questions

Determining Results

- We compare the annotations from our experiment with the final values used in the original research teams report (“original values”) and measure “agreement” with those values
  - 100% agreement unlikely – many questions require subjective judgement
  - Can’t measure “accuracy” directly – no known true values to compare
- Original values based on multi-step process: initial annotation by multiple different annotators, multiple review phases, etc.
  - Therefore, the original values provide a reasonable proxy for accuracy
- NOTE: For about 20% of the questions, we were able to find an answer for question original team marked as “unknown”
  - Further work will consider how to best evaluate these, excluded from current analysis
  - Could indicate even higher accuracy, or could indicate hallucinations

Research Questions

- Does utilizing the AI tool save time compared with manual annotation?
- Are annotators less accurate when using an AI tool, and if so by how much?
- Are there any patterns to accuracy/time savings in the types of questions/tasks that can help inform other researchers considering using AI tools to assist with data annotation and information extraction tasks?

Results

Mean Time per Question

Overall

Condition	Time per Question	Sample Size (n)
Control (Human Only)	4.90 minutes	91
Treatment (AI Tool)	**0.54 minutes*** **	106
Time Savings	**89%**	

By Annotator

	Annotator A	Annotator B
Control	5.28 minutes	4.00 minutes
Treatment	**0.50 minutes*** **	**0.625 minutes*** **
Time Savings	**91%**	**84%**
Sample Size (n)	134	63

Key Findings:

- As expected, using the AI tool produced strongly significant time savings
- A nearly 10x reduction in the amount of time taken to complete each question
- Some difference between the two annotators
- Subjectively, annotators reported completing the task with the tool assistance as being much more "pleasant"

* p<0.10, ** p<0.05, *** p<0.01. Mann-Whitney U test, control vs treatment. Time recorded to the nearest 30s.

Time Taken for Annotation (Per Question)

By Question Complexity

Condition	Low	Medium	High
Control	4.31 minutes	5.00 minutes	5.06 minutes
Treatment	**0.50 minutes*** **	**0.53 minutes*** **	**0.58 minutes*** **
Time Savings	**89%**	**89%**	**89%**
Sample Size (n)	40	85	72

Key Findings:

- More complex questions took somewhat longer using both the control and treatment methods, but the relative time savings remained consistent.

* p<0.10, ** p<0.05, *** p<0.01. Mann-Whitney U test, control vs treatment. Time recorded to the nearest 30s.

Accuracy (Agreement with "Original Values")

Overall

Condition	Accuracy	Sample Size (n)
Baseline (Random Guess)	27.5%	197
Control (Human Only)	43.9%	91
Treatment (AI Tool)	**61.3%****	106

By Annotator

Condition	Annotator A	Annotator B
Control	50.0%	29.6%
Treatment	**57.1%**	**69.4%*** **
Sample Size (n)	134	63

* p<0.10, ** p<0.05, *** p<0.01. Fisher's exact test, control vs treatment

Key Findings:

- Using the tool significantly **improves** accuracy, an unexpected result
- Annotator A had some subject matter expertise with the task/domain, but little experience using AI tools. Annotator B was the opposite, having less experience with the task/domain, but considerable experience using AI tools
- Annotator skill level had a much larger impact than expected. Annotator B, despite performing much worse at the control task, performed considerably better when using the tool.

Accuracy (Agreement with "Original Values")

By Question Complexity

Condition	Low	Medium	High
Control	56.2%	50.0%	30.3%
Treatment	**83.3%***	**72.1%****	**35.9%**
Sample Size (n)	40	85	72

* p<0.10, ** p<0.05, *** p<0.01. Fisher's exact test

Key Findings:

- Using the tool significantly improved accuracy for low/medium complexity question, but not for the most complex questions.
- Accuracy was low for both the control and treatment groups on high complexity questions.
- These questions tended to require the most subjective judgement, and could be explained by more limited contact with the original research team or lower annotator skill
- However, these questions also often required referring to many different portions of the documents and synthesizing complex conclusions. This may indicate a limitation of the fairly basic RAG approach used

Accuracy (Agreement with "Original Values")

Low/Med Complexity Questions Only

Condition	Annotator A	Annotator B
Control	54.5%	42.9%
Treatment	**76.2%****	**76.0%***
Sample Size (n)	86	39

High Complexity Questions Only

Condition	Annotator A	Annotator B
Control	40.0%	15.4%
Treatment	**28.6%**	**54.5%***
Sample Size (n)	48	24

Key Findings:

- Given the limited sample size, difficult to draw strong conclusions about the high complexity questions
- Annotator A performed worse with access to the tool on high complexity questions (though not statistically significant), while Annotator B performed significantly better.
- This could indicate that skill using AI tools is especially important on more complex data annotations (vs not as important on less complex ones)

* p<0.10, ** p<0.05, *** p<0.01. Fisher's exact test

Conclusions

- In highly complex, challenging data annotation tasks, even a basic RAG approach was able to substantially **improve** the performance of untrained human annotators, an unexpected result
- Allowing human annotators to use AI tools reduced task time by a factor of 10x
  - This could allow researchers to assemble data annotations for much larger datasets than would otherwise be feasible
- We find some evidence that annotator skill/experience with AI tools is a key factor in their effectiveness
  - Researchers should consider incorporating training for annotators if they intend to utilize these tools
- We find some evidence tool use may be less effective on extremely high complexity tasks, but further research is needed in this area

Limitations/Further Work

- Accuracy comparison is limited due to no known true values
  - Further work should include a detailed qualitative review of a sample of non-agreement questions to attempt to discern true values
- Limited sample size
  - Human annotation is cost prohibitive, the motivation for this research
- RAG approach used is simplistic, and only 1-2 prompts used per question
  - Further work should evaluate newer and more robust approaches
- Line between AI and human judgement is blurred
  - Further work should include a 100% AI approach as comparison
  - Further study is needed for the effect of annotator skill with AI tools

References

- Beltran, Daniel O., Hannah Bensen, Amy Kvien, Erin McDevitt, Monica V. Sanz, and Pinar Uysal (2023). “What are Large Global Banks Doing About Climate Change?,” International Finance Discussion Papers 1368. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/IFDP.2023.1368.
- Es, Shahul, Jithin James, Luis Espinosa-Anke, Steven Schockaert (2023). “RAGAS: Automated Evaluation of Retrieval Augmented Generation,” arXiv. https://doi.org/10.48550/arXiv.2309.15217
- Kamradt, Greg (2023). “Needle in a Haystack Analysis,” X, https://x.com/GregKamradt/status/1722386725635580292