Zero-Shot Information Extraction with Community-Fine-Tuned Large Language Models From Open-Ended Interview Transcripts

Nazmul Kazi[1], Indika Kahanda[2], S. Indu Rupassara[3] and John W. Kindt Jr.[4]

[1,2] University of North Florida, Jacksonville, FL, USA. [3,4] FruitVaccine, Inc., Champaign, IL, USA.

[1]nazmulkazi@oxiago.com [2]indika.kahanda@unf.edu [3]indurupassara@gmail.com [4]jowaki@comcast.net

Kahanda BioMedInfo Lab

Introduction

- Surveying and interviewing are common methods for gathering information from customers or stakeholders.
- Extracting information from long interview transcripts is a time-consuming, resource-intensive, and tedious process.
- Text mining offers automation and superior efficiency over manual labor for large datasets or long-term projects.
- Model training is resource-intensive and raises challenges for medium-sized datasets or one-time extractions.
- We propose a hybrid, human in the loop, approach utilizing Community-Fine-Tuned Large Language Models (CLLMs) for rapid data extraction.
- Our project aims at one-time data extraction, excluding data annotation and LLM fine-tuning.
- The ML pipeline presented here was developed in late 2021 before the release of ChatGPT in November 2022.

Dataset

- In 2021, FruitVaccine, Inc. interviewed 135 key individuals from vaccine-related industries to assess the market for oral vaccines [1].
- They asked 29 questions during the interviews. We narrowed our focus to seven questions for this publication.
- We leveraged Google Cloud Platform Speech-to-Text to transcribe the first 27 interviews and Zoom's native speech recognition technology for the remaining interviews.

Interview Questions:

1. What is your role as it relates to vaccines?
2. Can you tell us about your typical workflow?
3. How long does the vaccination process take per patient?
4. If you could improve one step or thing in your workflow process, what would it be?
5. What percentage of patients are afraid of needles?
6. Do you have any unmet needs regarding the vaccination process? If yes, what are those?
7. Can we keep your contact information?

Approach

- Our initial goal was to extract exact answers to various interview questions.
- Analyzing long transcripts is challenging since they exceed the token limitations of CLLMs. Besides, similar questions were asked during the interviews.
- We broke down the main task into three subtasks and employed a CLLM for each.
- We picked the best-performing CLLMs based on their performance reported on benchmark datasets.

Sub-task	Utilized CLLM
Semantic Text Matching	SRoBERTa[a]
Exact Answer Extraction	RoBERTa-base for QA (2nd rev)[b]
Sentiment Analysis	RoBERTa Sentiment[c]

[a]huggingface.co/sentence-transformers/paraphrase-distilroberta-base-v2
[b]huggingface.co/deepset/roberta-base-squad2
[c]huggingface.co/siebert/sentiment-roberta-large-english

Overall Pipeline

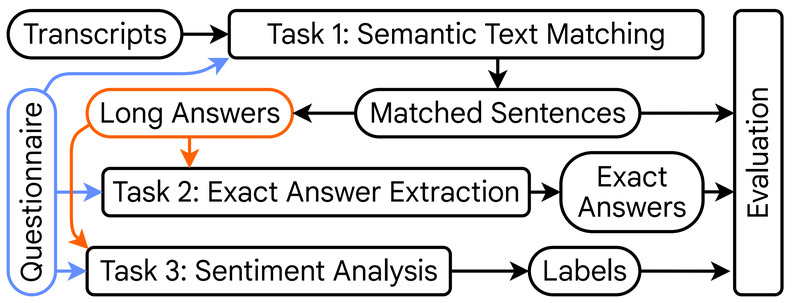
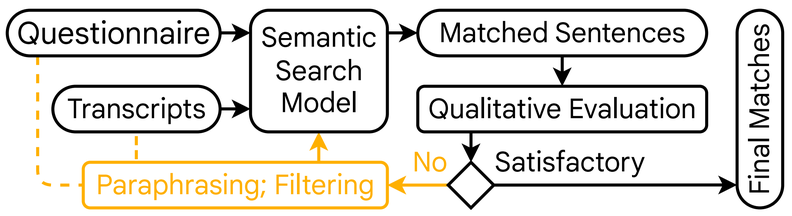
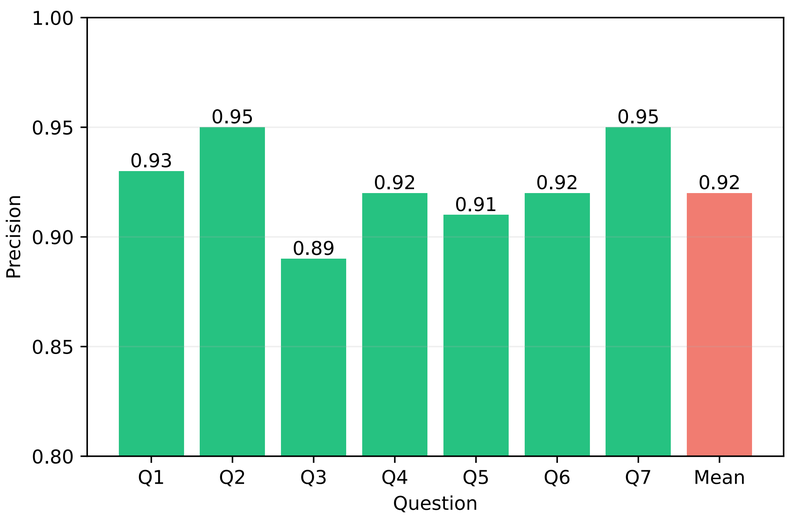


Figure 1: An overview of our approach for extracting information from survey-interview responses in digital transcripts using CLLMs.

Semantic Text Matching (STM)

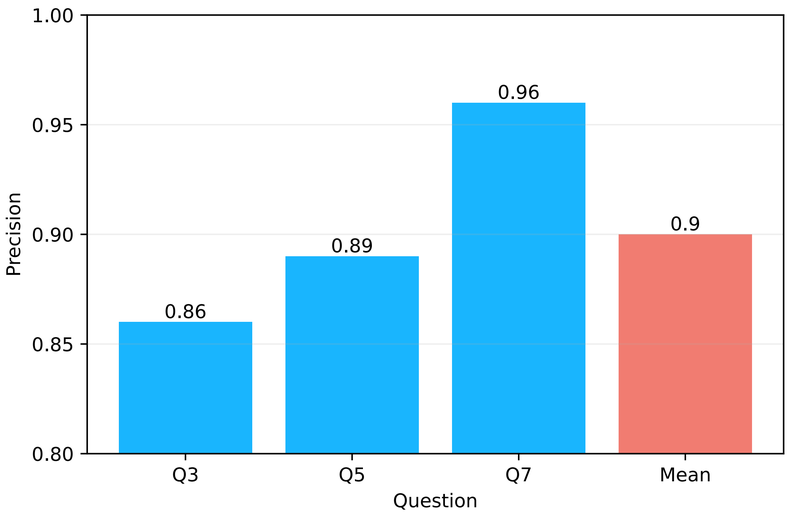


- Identifying questions in the transcripts to narrow down the context for our EAE and SA models.
- Evaluated model outputs qualitatively for satisfactory performance rather than peak performance. Paraphrasing and filtering are employed to improve performance.



Exact Answer Extraction (EAE)

- Text between two questions is regarded as a long answer.
- Answers that seem out of scope are considered incorrect.



Sentiment Analysis (SA)

- Discerning majority sentiment (positive/negative) of responses to a specific question.
- We pass the question as context. Two distinct questions were used to interview and we experimented with both.

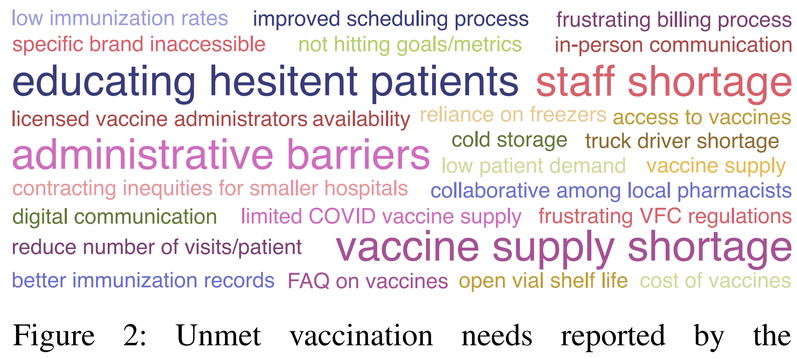
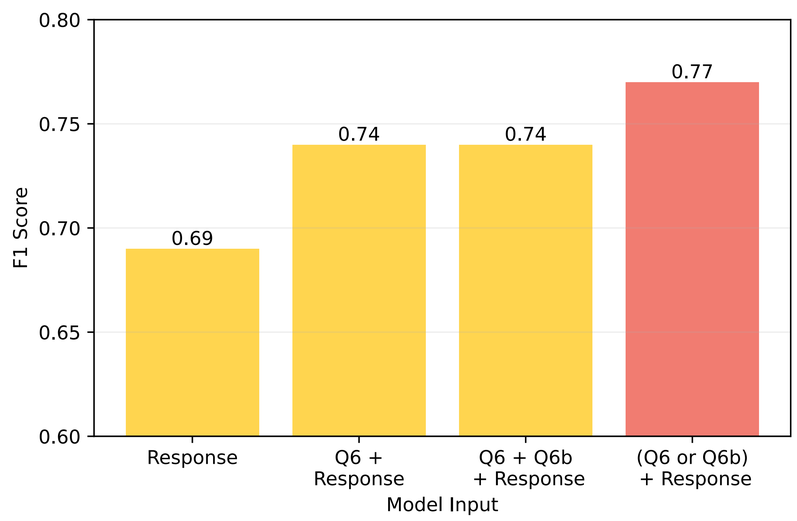


Figure 2: Unmet vaccination needs reported by the interviewees, coded into short terms for generalization. These terms are not input to the model, but rather a visual representation of the response nature only.



Conclusions and Future Work

- We utilized CLLMs in three distinct survey text data analysis tasks.
- Hybrid semi-automated pipeline combines CLLMs' rapid information extraction with manual curation for high-quality insights from survey responses.
- STM and EAE models showcase impressive performance, highlighting the potential of CLLMs in streamlining text data analysis without fine-tuning.
- Our approach can be extended to unexplored tasks, such as image classification, fill mask, summarization, etc.
- Ensembling multiple models for the same task can be investigated for improved output quality.

Acknowledgements

This Customer Discovery Survey was funded by the NSF-SBIR (Award No. 2026281) and the Supplemental I-Corps Award received by FruitVaccine, Inc. The computations presented here were conducted in the National Research Platform's Nautilus HyperCluster.

References

[1] S. Indu Rupassara, John W. Kindt Jr, Nazmul Kazi, and Indika Kahanda. Challenges and opportunities in current vaccine technology and administration: A comprehensive survey examining oral vaccine potential in the united states. *Human Vaccines & Immunotherapeutics*, 0(0):2114422, 2022. doi: 10.1080/21645515.2022.2114422. PMID: 36082816.