MACHINE LEARNING

CS5821

PROJECT REPORT ON

META - EVALUATION CHATBOT

Enhancing Accuracy and Transparency in Evaluation Processes

Team Mavericks
Arun Totad
Visha Shende
Akshara Reddy
Shivani Rana
Swarna Doppa

Under guidance of:
Dr. Alvis Fong
Zach Tilton (Evaluation specialist)

ABSTRACT

This project spearheads the development of a novel Large Language Models (LLMs)-powered chatbot for meta-evaluation. It leverages the capabilities of LLM's to analyze and summarize information from uploaded PDFs. By integrating advanced Natural Language Processing (NLP), the chatbot achieves a high degree of context awareness, enabling users to ask intricate questions about the document's content. Primarily designed with a specific evaluation domain in mind, the system's modular architecture allows for future adaptation to diverse industries through parameter optimization. This expands the chatbot's potential applications beyond its initial scope. The technical foundation relies on the Llama2 LLM model, integrated using Retrieval Augmented Generation (RAG) techniques. Additionally, Chain Lit facilitates deployment, while a Chroma vector store enables context-aware information retrieval. This synergy empowers the chatbot to effectively process queries and provide relevant summaries. This project contributes to the advancement in LLM-implemented applications for meta-evaluation. Demonstrating the effectiveness of NLP and context-aware processing, it paves the way for more sophisticated tools to assist human evaluators. The modular design allows for future expansion and customization, ultimately aiming to enhance accuracy and efficiency in meta-evaluation processes.

Apr 18, 2024

TABLE OF CONTENTS

INTRODUCTION

1.1. Project Overview

The ever-growing field of meta-evaluation seeks to assess the quality and effectiveness of existing evaluations. This process is crucial for ensuring the trustworthiness and usefulness of evaluation findings in various domains. Traditionally, meta-evaluation has relied on manual methods, which can be time-consuming and labor-intensive. However, recent advancements in Large Language Models (LLMs) offer exciting possibilities for streamlining and enhancing this process. This project delves into the development of a novel LLM-powered chatbot specifically designed to assist human evaluators in meta-evaluation tasks. This innovative tool leverages the power of LLMs to analyze and summarize information extracted from uploaded PDF documents. By incorporating advanced Natural Language Processing (NLP) techniques, the chatbot attains a high degree of context awareness, enabling users to ask intricate questions about the content of the documents. Furthermore, the project adheres to the established 30 Meta-Evaluation Standards, ensuring a comprehensive and ethically sound approach to meta-evaluation. By integrating these Standards into the development process, the chatbot strives to promote responsible and transparent use of evaluation findings. This report outlines the functionalities, technical foundation, and potential significance of this LLM-powered chatbot for meta-evaluation. It also explores the project's results and its potential to revolutionize the way meta-evaluation is conducted.

1.2. 30 Meta Evaluation Standards

The development of this LLM-powered chatbot prioritizes upholding the established 30 Meta-Evaluation Standards. These standards, outlined by the Joint Committee on Program Evaluation Standards (JCSEE), serve as a comprehensive framework for ensuring the rigor, credibility, and ethical conduct of evaluations (Stufflebeam, 2016, 7). By adhering to these standards throughout the chatbot's development and implementation, we aim to:

- **Enhance Utility (U Standards):** The chatbot is designed to cater to the specific needs of evaluators by providing relevant information extracted from uploaded documents (U5). It facilitates clear communication between stakeholders throughout the meta-evaluation process (U1).
- **Ensure Feasibility (F Standards):** The system is designed to be user-friendly and accessible to evaluators with varying levels of technical expertise (F2). Additionally, the modular architecture allows for cost-effective adaptation to diverse evaluation needs (F3).
- **Maintain Propriety (P Standards):** The project emphasizes clear identification and ongoing engagement with all stakeholders involved in the evaluation process (P1). This includes program implementers, beneficiaries, funders, and of course, the primary evaluators themselves.
- **Guarantee Accuracy (A Standards):** The chatbot leverages NLP techniques to ensure accurate information extraction from documents (A1). The system also promotes transparency by allowing users to understand the rationale behind the chatbot's summaries and responses (A9).

Throughout the development process, particular emphasis was placed on adhering to the **Accuracy Standards (A3: Negotiated Purposes)**. Additionally, the project upholds the **Accuracy Standard (A6: Meaningful Processes and Products)** by prioritizing the generation of clear, concise, and interpretable summaries of the extracted information.

By adhering to these core principles outlined in the 30 Meta-Evaluation Standards, this LLM-powered chatbot strives to contribute to a more efficient, accurate, and ethically sound approach to meta-evaluation.

1.3. Objectives

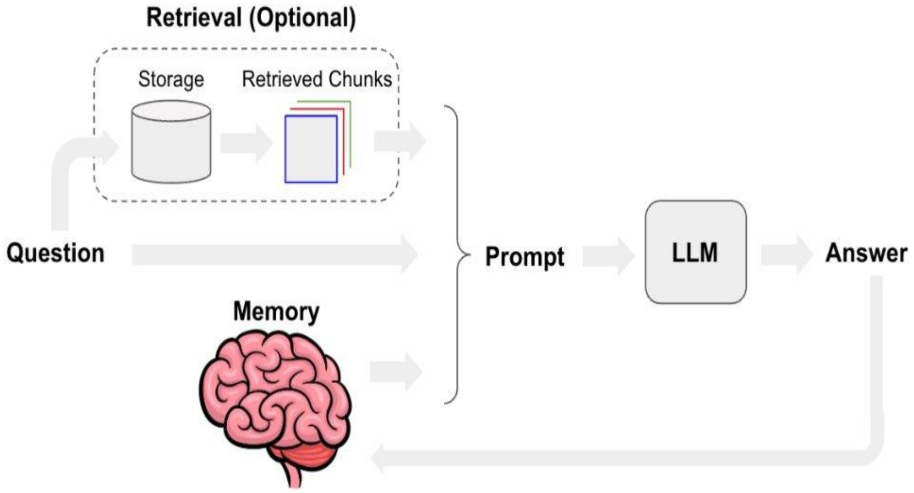
- Develop an LLM-powered chatbot specifically designed to assist human evaluators in meta-evaluation tasks.
- Leverage the capabilities of LLMs to analyze and derive insights from uploaded PDF documents.
- Integrate advanced Natural Language Processing (NLP) techniques to achieve a high degree of context awareness so we could analyze the evaluation standards with the most accuracy.
- Ensure adherence to the established Meta-Evaluation Standards, there are 30 however we focused primarily on Accuracy for this project.
- Design a system that is adaptable and modular, allowing for future expansion and customization to cater to the specific needs of diverse evaluation scenarios.

METHODOLOGY

The methodology section of this report outlines the approaches, techniques, and tools employed in the development and evaluation of the project. Each methodology plays a crucial role in achieving the project objectives by providing systematic frameworks for data processing, analysis, and interpretation.

2.1. RAG

Retrieval-augmented generation (RAG) is a technique for enhancing the accuracy and reliability of generative AI models with facts fetched from external sources. In other words, it fills a gap in how LLMs work. Under the hood, LLMs are neural networks, typically measured by how many parameters they contain. An LLM's parameters essentially represent the general patterns of how humans use words to form sentences. That deep understanding, sometimes called parameterized knowledge, makes LLMs useful in responding to general prompts at light speed. However, it does not serve users who want a deeper dive into a current or more specific topic (Merritt, 2023).



2.2. OLLAMA

Optimized Language Learning and Acquisition Methodology (OLLAMA) focuses on enhancing language learning processes through various optimization techniques. These techniques may include spaced repetition, contextual learning, and adaptive feedback mechanisms. By leveraging these strategies, OLLAMA aims to improve language acquisition efficiency and effectiveness, ultimately facilitating more rapid and comprehensive language learning outcomes.

2.3. Lang Chain

Lang Chain is a systematic approach to language processing and analysis, specifically tailored for tasks involving natural language understanding. This methodology involves chaining together linguistic elements such as words, phrases, and syntactic structures to infer meaning and context from textual data. By analyzing the sequential relationships between language components, Lang Chain enables robust comprehension and interpretation of textual content, making it particularly valuable in applications such as text summarization and sentiment analysis (Langchain, n.d.).

2.4. Chain-Lit

Chain-lit methodology focuses on the extraction and analysis of chained literals within textual data. Chained literals refer to sequences of related terms or concepts that are implicitly connected within the text. By identifying and analyzing these chains of literals, Chain-Lit can reveal underlying semantic structures and associations, providing valuable insights into the content and context of the text. This methodology is especially useful in tasks such as information extraction, semantic analysis, and knowledge discovery from textual data (*LangChain*, n.d.)

SYSTEM FUNCTIONALITY

3.1. Data Input

Meta evaluation starts when users interact with the chatbot by uploading a PDF of the primary evaluation which needs to be assessed. Upon initiation of the chat session, users are prompted to upload a PDF file. Users are guided through the upload process, ensuring seamless interaction. The challenge lies in the fact that processing large PDF files consumes considerable time when executed locally.

3.2. Pre-processing

Uploaded PDF documents undergo pre-processing to prepare them for analysis. This includes extracting text from the PDF using PyPDF2 library, splitting the text into manageable chunks, and converting it into a format suitable for natural language processing (NLP). The RecursiveCharacterTextSplitter object divides the text into smaller chunks, optimizing it for further processing by the language model.

3.3. Context-Aware Processing

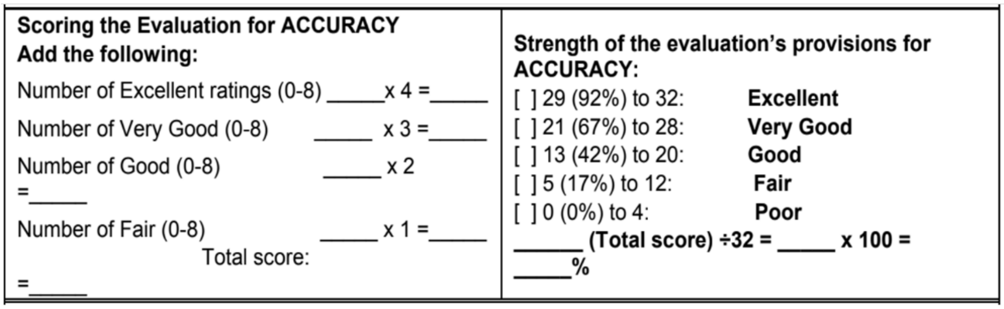
Utilizing natural language processing (NLP) techniques, the chatbot extracts relevant information from the pre-processed text. This involves identifying key concepts, entities, and relationships within the document. The Ollama embeddings model is employed to capture semantic information and facilitate effective information retrieval. The chatbot employs context-aware processing to understand user queries in the context of the uploaded document. This involves semantic matching, understanding user intent, and considering the ongoing conversation to provide relevant and accurate responses.

3.4. Meta-evaluation standards

There are 30 meta-evaluation standards of different categories and as we are only focusing on Accuracy for the scope of this project, we will only consider the 8 meta-evaluation standards that correspond to it. Each of these 8 meta-evaluation standards have 6 checkpoints in them. Each checkpoint is posed as a question to the bot and the questions are posed in a way that results in a True or False answer.

3.5. Scoring technique

The questions(checkpoints) are pre coded in the code since we do not want the user to bother with entering the question manually. The true and false answers are then scored with 0 and 1 and we calculate grade for meta evaluation standards. So there will be ideally 8 grades for accuracy alone. Here we had a challenge because the checkpoints were rather subjective and the bot could only answer the first couple of sub standards, other were quite ambiguous for True and False answers. So we can say around 12 checkpoints or questions. Once we have the score, we identify individual grades and then finally the grades are fed with a mathematical formula (snapshot on right) to calculate the finalized score.



Scoring the Evaluation for ACCURACY
Add the following:

Number of Excellent ratings (0-8) _____x 4 =_____

Number of Very Good (0-8) _____ x 3 =_____

Number of Good (0-8) _____ x 2 =_____

Number of Fair (0-8) _____ x 1 =_____

Total score: =_____

Strength of the evaluation's provisions for ACCURACY:

[ ] 29 (92%) to 32: **Excellent**
[ ] 21 (67%) to 28: **Very Good**
[ ] 13 (42%) to 20: **Good**
[ ] 5 (17%) to 12: **Fair**
[ ] 0 (0%) to 4: **Poor**

______ (Total score) ÷32 = _____ x 100 = _____%

3.6. Output

The chatbot presents information to users in a clear and concise manner. This includes response in True and False form for all the pre coded questions, description of how the BOT arrived at that conclusion, and any relevant sources or references.

KEY FINDINGS AND CHALLENGES

3.7. Findings

This project yielded several noteworthy insights for the entire team. Through the integration of Lang-Chain, RAG, OLLAMA, and Chain-Lit, we were able to understand many aspects of Large Language Models and their implementation techniques. The key findings are as follows:

- **Consistency in Evaluation Criteria**: We were not able to test it on many of the primary evaluations because the business was unable to provide us with test data during this project. However, we have a real meta-evaluation standard which we could reduce to create a primary evaluation document and the testing was done with that. A significant trend was observed in the evaluation technique of this application that modifying the sentences may change the way the response is given. This suggests a foundational understanding and great importance on how a particular subject can be perceived.
- **Quality of Standards:** The meta evaluation standards revealed varying degrees of guidelines. While some standards excelled in clarity and completeness of the checks it should perform, others exhibited subjectivity in response particularly in Utility, feasibility, and Accountability criteria.
- **Integration of Stakeholder Perspectives:** Notably, the incorporation of business stakeholder perspectives emerged as a strength during this project. This inclusivity contributed to a more comprehensive understanding of the business use case and the evaluated subject matter.
- **Sources and reasoning:** This chatbot provides us with its reasoning behind giving a particular response proved to be useful for the business users to understand the basis of the scores we provide. This also helped us understand when the reasoning from the model is flawed and needs improvisation.
- **Areas for Enhancement:** Lastly, the meta-evaluation identified common areas for enhancement across this project, including the need for enhanced transparency in methodologies, better modifications in the model parameter to ensure accurate results, increased rigor in data interpretation, and improved alignment with subjective evaluation objectives.

3.8. Challenges

- When the size of primary evaluation was big, there was a considerable amount of time spent on the uploading of the information for the BOT. We anticipate the reason to be the execution in, local machine.
- Looking at the checkpoints we can say that many of the criteria are subjective and difficult for our process to return as a true or false answer. It can impact our scoring mechanism. One way to tackle this would be to feed the response of BOT to another sentiment analysis process and if the feedback is positive, we can accept the response as True however that may have incorrect response.
- We realized that reframing the questions leads to a different response from the BOT. Even subtle changes can cause the model to change the way it handles and returns the answer. This makes this process very susceptible to changes.
- Since we have coded the checkpoint for the meta-evaluation standards this code is only useful for the primary evaluations which can be reevaluated on these standards. This process is not intelligent enough yet to find the standards itself and grade the document dynamically.

CONCLUSION

Despite encountering challenges with access to primary evaluation data during this project, the meta-evaluation using a real meta-evaluation standard provided valuable insights into evaluation techniques and standards. The analysis underscored the foundational importance of how language and sentence structure can influence perception, highlighting the need for careful consideration in crafting evaluation criteria.

The evaluation of standards revealed a spectrum of guideline effectiveness, particularly in the areas of clarity, completeness, and subjectivity in response. This underscores the importance of refining standards, especially in aspects related to utility, feasibility, and accountability criteria, to ensure robust and objective evaluations.

An exceptional aspect of this project was the integration of business stakeholder perspectives, enriching the understanding of both the business use case and the evaluated subject matter. The engagement of stakeholders contributed significantly to the project's comprehensiveness and relevance.

Moreover, the utilization of reasoning provided by the chatbot proved invaluable, offering transparency in the evaluation process and aiding in identifying areas for model improvement. This approach facilitated enhanced collaboration between data science and business stakeholders, ensuring clarity and trust in the evaluation outcomes.

Looking ahead, the meta-evaluation highlighted key areas for enhancement within the project, emphasizing the need for improved transparency in methodologies, more precise modifications to model parameters for accuracy, rigorous data interpretation, and better alignment with subjective evaluation objectives. These insights will guide future endeavors, fostering continuous improvement in evaluation practices and standards.

REFERENCES

1. Langchain. (n.d.). Introduction | 🦜 LangChain., from https://python.langchain.com/docs/get_started/introduction , Retrieved April 18, 2024
2. Merritt, R. (2023, November 15). *What Is Retrieval-Augmented Generation aka RAG?* NVIDIA Blog., from https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/ , Retrieved April 18, 2024
3. Stufflebeam, D. L. (2016). *PROGRAM EVALUATIONS METAEVALUATION CHECKLIST*. Daniel L. Stufflebeam, from https://drive.google.com/drive/folders/1sy0kGC_79PkxLgzWzA9oXjm_HLV_Tzqt , Retrieved March 2024
4. *LangChain*. (n.d.). Chainlit. Retrieved April 18, 2024, from https://docs.chainlit.io/integrations/langchain