DARIA-3o: Chatbot for InfoTunnel

Team 14
Saivarun Tanjore Raghavendra
Mohammed Tareq Sajjad Ali
Suraj Poldas
Mano Harsha Sappa

AIT 526 - Introduction to NLP

Under the Guidance of

Dr. Lindi Liao

George Mason University

Abstract:

This project showcases the custom framework for a chatbot created for InfoTunnel, an integrated platform for information management. This chatbot provides the best user experience and helps locate any information on the website in an effective and intuitive way without necessarily having to sift through them physically. Leveraging state-of-the-art natural language processing models, including Microsoft's Phi-2 and Meta's LLaMA-2 from Hugging Face, the chatbot responds with answers that are accurate and contextually relevant to the user's queries. The system uses FAISS for efficient semantic similarity searches, while embeddings are done via Hugging Face to vectorize data for fast and accurate information retrieval.Besides, it has multimodal interaction-availability of inputting users' queries either in text or voice form and ease of interface for smooth user interaction. This project demonstrates the application of various state-of-the-art AI technologies in facilitating information retrieval on InfoTunnel and lays the groundwork for future expansions into other domains and platforms.

Keywords:

Chatbot Development, Large Language Models (LLMs), Natural Language Processing, Microsoft Phi-2, Meta LLaMA-2, Semantic Search, FAISS, Hugging Face Embeddings, Multimodal Interaction, Information Retrieval Systems.

Introduction:

In the modern-day digital world, finding specific information online has become akin to searching for a needle in a haystack. InfoTunnel is faced with this issue because copious amounts of data are organized on it, and users usually cannot find what they exactly need out of the extreme content provided.Classic search features return wide, less relevant results, which take the user to sift through information that is not needed. The solution will be proposed by developing a customized chatbot, which will make retrieving information from InfoTunnel faster, easier, and more accurately progressive.

The main contribution of this project is a chatbot that uses advanced models of natural language processing, including Microsoft's Phi-2 and Meta's LLaMA-2 from Hugging Face. These models will enable the chatbot to understand user queries and answer them with a lot of contextually relevant information.FAISS- Facebook AI Similarity Search-is used in the system for semantic similarity searches efficiently. Besides, Hugging Face embeddings are used to convert the website data into vectors, which makes the retrieval process fast and precise. The chatbot will be designed to support multimodal interaction, meaning users can input queries either through text or voice for a smooth and engaging user experience. This project will be proof of how AI can be practically used in order to enhance information access. Moreover, this sets the baseline for further extension into other platforms and domains in the near future.

- **Advanced NLP Models:** The integration of Microsoft's Phi-2 and Meta's LLaMA-2 in the working of the chatbot helps it give very accurate and relevant responses to user queries.
- **Semantic Search Efficiently:** Using FAISS for really fast similarity searches, this system can surface specific information from its vast data repository with accuracy.
- **Multimodal Interaction:** A user can interact through textual input and voice input, facilitating much easier access and therefore improving usability.
- **Embeddings and Vectorization:** This allows the chatbot, using Hugging Face embeddings, to efficiently process and index very large datasets.
- **Improved User Experience:** The chatbot provides a frictionless interface to the user by reducing cognitive load and allows him to concentrate on his work rather than navigating through the complex data structure.

System Architecture

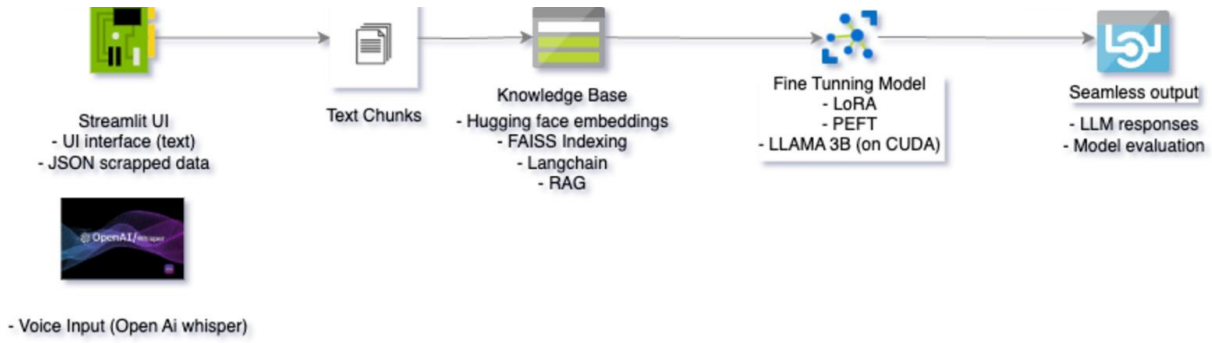


Figure1: Language Model Pipeline

It has quite a sophisticated method for the processing and generating of text answers. Semantic representation will be performed by Hugging Face embeddings, and FAISS indexing for efficiently getting the needed pieces of information in a huge knowledge base. Thus, it produces responses via a fine-tuned language model, most probably based on LLaMA 3B. The given LLaMA 3B model is then further adapted using some other techniques like LORA and PEFT that allow efficient fine-tuning. Besides, it allows the use of text and voice inputs, making the system quite versatile and user-friendly. While the system is promising, its real-world performance has to be tested further for evaluation and subsequent improvements.

Proposed Methods

1. User Interface and Input Device Management:

The user interface for the chatbot is made using the Streamlit framework, which will serve as an intermediate for communicating with the chatbot. In fact, Streamlit is good at creating web-based applications where the chatbot will be deployed and maintained without much effort. The users will have several options like text input, where they enter their queries directly. This will be a simple interface that lets the chatbot process a user's query and respond in the required manner quickly. Text input is preferred by people who don't want to speak and is also the most orthodox form of communication in web applications.

Apart from text input, the system is also having a voice input feature taken care of by the OpenAI's whisper model to transcribe the voice into text. In other words, people can also speak their queries instead of typing them, thus further increasing the ease of accessibility to the chatbot enhancement in one's overall experience. Voice input is very useful for people on the move or by those for whom speaking is easier than typing. The Whisper Model is highly robust in speech recognition, which would ensure high-quality transcriptions in even noisy surroundings; thus improving the chatbot's reliability in real-life use cases.

2. Loading and Processing Content

The loading and processing portion of the content is made possible through a JSON content loader. This loader extracts and processes contents such as those relevant to the entity in JSON files. Such a file contains user-generated data or requests that a user may have directed at the chatbot. The JSON loader ensures that the information is retrieved and parsed easily within the structure, providing a structured means for the system to handle it. This step would be extensive given that it allows transforming the raw data into a usable one that could be searched, analyzed, and responded to by the system-all critical elements that affect the performance of the imperfect chatbot.

The text that has undergone content loading is text splitting, which is the process of chunking large blocks of text into smaller, more manageable pieces. It is primarily intended for ensuring that such chunks are used effectively with respect to their meaning: so that the chatbot can process and understand what they contain. Smaller but context-based bits of information are constructed for easier analysis and retrieval when needed; furthermore, this step gives the system more great capability for overcoming huge amounts of data because it relieves it of parts of its computational load and speeds up its performance at the same time.

3. Knowledge Base Management

Embedding Generation is the major role player in transforming textual chunks into numerical representations for effective processing by the chatbot. It uses Hugging Face embeddings, thus vectorizing the contents into high-dimensional vectors. These vectors are responsible for capturing the semantics meaning of the text and hence able to compare up with contents at the deeper level in terms of meaning. Embedding is a critical aspect of natural language understanding, as it enables the chatbot to recognize some relationships between the data and bring up the most relevant one concerning the user's query.

Once the embeddings are generated, they are stored in FAISS Knowledge Base. The FAISS (Facebook AI Similarity Search) is one powerful tool for indexing and similarity searches over large datasets. The embeddings are indexed within FAISS, which allows the fast retrieval of the most relevant contents in response to user queries. FAISS has been optimized to ensure that search efficiency enables even the chatbot to use large volumes of data without sacrificing speed and accuracy. Thus, the entire system ensures scalability and performance as the knowledge base grows.

4.. Model fine-tuning and integration.

Fine-tuning is done on the chatbot language models by preprocessing datasets using advanced transformer libraries. Fine-tuning makes a model suitable for a particular domain and even the requirements of the chatbot. This adaptation is also pivotal for attaining much-needed specificity in the accuracy of the responses. Learning the InfoTunnel dataset nuances is very important since fine-tuning is a prerequisite for more contextually relevant and accurate responses from the chatbot.

The efficiency with which the fine-tuning process is facilitated has included LuoRA adaptation with PEFT techniques, enabling the model's layer update functionality of the language models at a rapid pace without requiring the exertion of many computational resources. With these adapting models, already fine-tuned for this application, Meta LLaMA 3-8B, and Microsoft Phi 3.5 have increased capacity with regard to response generation. The specialized knowledge and conversation depth added by these models make the chatbot more effective at addressing its users' queries.

5. Response Generation

In this phase of Response Generation, the chatbot initiates the process with a prompt constructed based on the context retrieved from the FAISS knowledge base. Such prompt serves as an input for the language models to understand the user's query based on the data stored in the database. The context is very important in generating accurate and contextually appropriate responses. In fact, retrieved context will be integrated into the prompt so that answers will be accurate and relevant to the specific user query.

Then, the system evaluates the responses produced by Meta LLaMA and Microsoft Phi models on several core stages: accuracy, relevance, completeness, and clarity. Evaluations of the chatbot will ensure the generation of high-reliability answers. The presence of evaluation checks determines whether the response is presented in a way that answers the user question. It will check completeness, clarity, and understanding of the answer. With great guarantees, all these factors judge the answer on top-notch quality as well as maximum user satisfaction.

6. Model Evaluation

The radar graph very clearly makes the comparative performance of Microsoft Phi-2 and Meta LLaMA apparent across five critical evaluative parameters: Contextual Understanding, Data Analysis, Comparative Analysis, Technical Explanation, and Practical Applications. Meta LLaMA shone brighter for Contextual Understanding and Technical Explanation, which is a proven ability to respond in nuanced, detailed, and precise contextual terms. Whereas, Michosoft

Phi-2 was a definite outperformer in Data Analysis and Practical Applications: a clear indication of its efficiency in retrieving and presenting structured, factual data.

This multimodal feature of the chatbot is perhaps one of the most exciting features of the system, requiring text responses and importing other text as visual images based on keyword searches. It gives an interesting user experience, being two forms of response candidate. Such an operation-the system is able to do this through the integration of JSON data search methods that search for keywords and retrieve images relevant to those keywords-isn't only improving user engagement but is increasing the system's applicability for various user scenarios.

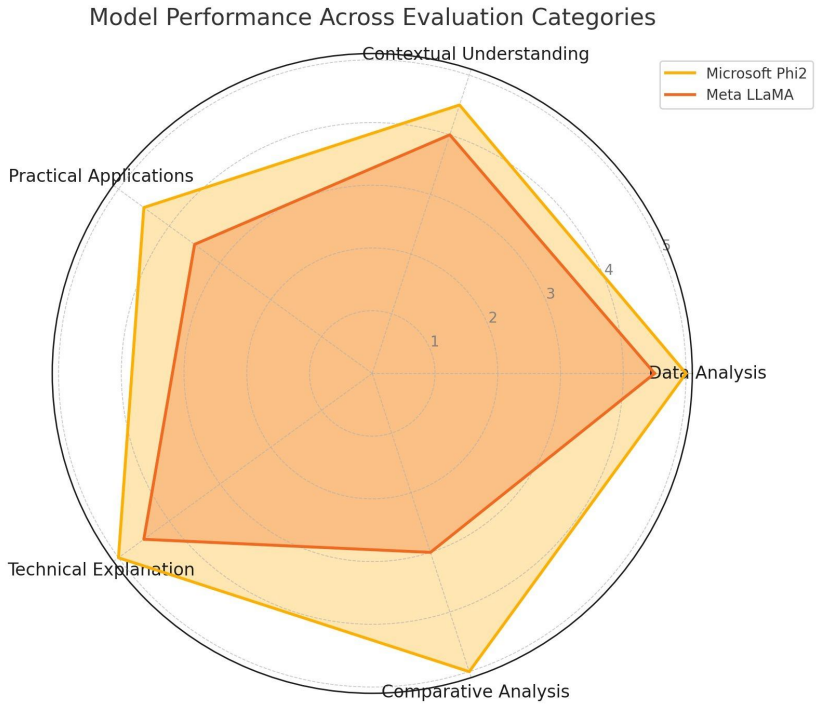


Figure2: Microsoft Phi-2 and Meta LLaMA, across five evaluation categories:

The radar chart provides a comparison of performances of Microsoft Phi-2 and Meta LLaMA in evaluation categories. The lines from the center represent the length on which the performance of the model i.e., its performance in that particular category. Clearly, the chart indicates the strengths and weaknesses of the two models: The first model is good in terms of Contextual Understanding and Practical Applications, while the second model excels better in Technical Explanation and Comparative Analysis. In the end, the choice depends on what the application requires in that regard.

Key Observations:

- Meta LLaMA not only tackles vague and contextually dependent questions but also very responsive to analytical tasks.
- Microsoft Phi-2 gives short and data-rich answers but excels with numerical and fact-oriented queries.
- Multimodal features boost the enjoyable experience for the user by generating both text and image outputs-they are supposed to interact better or more engagingly-in.
- It shows openings that can balance the strengths of the model towards a more holistic response generation system.

The bar graph shows how the average points were distributed among different evaluation criteria for both Microsoft Phi-2 and Meta LLaMA models - wrapping the strengths and weaknesses around these two models. The considered evaluation criteria include Data Analysis, Contextual Understanding, Practical Applications, Technical Explanation, and Comparative Analysis. All these evaluation areas essentially capture the ability to cater to different facets of user queries by the models.

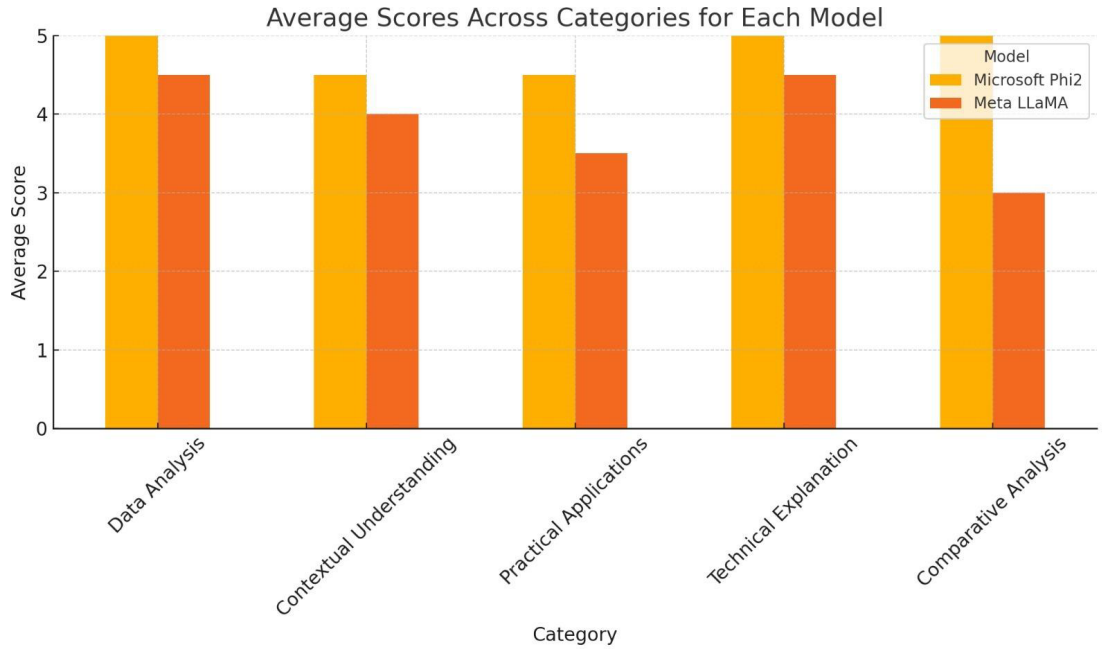


Figure3: Model Performance Comparison: Microsoft Phi-2 vs. Meta LLaMA

Microsoft Phi-2 has the best average scores in Data Analysis and Practical Applications. The reason it stands out has to do with converting structured data into concise actionable insights and thus being suited to heavy data processing. Also, the highest performance can be seen through dealing with repeated user queries that are too often very specific in terms of factual matters.

Meta LLaMA is the best performer in Technical Explanation, Comparative Analysis, and Contextual Understanding, as this specific model possesses much more nuanced, context-aware approaches toward ambiguous and exploratory problems. Another area for differences in user outputs would be open-ended questions and a need for elaborate comparison, like in the first two cases, where Meta would give a richer analytical answer than Microsoft could offer.

Insightful deductions:

- Analyzing data: Microsoft Phi-2 is an excellent one that excels in retrieving data points accurately and factually for those applications that need to deliver precisely the required outcome.
- Understanding the context: Meta LLaMA brings more heat to the situation when it comes to comprehending and responding to tougher or more ambiguous questions in offering more context and depth in their answers.
- Wielding practicalities: Both perform well, but Microsoft outpaces slightly against the tests in term of efficiency with which it gives concise actionables.
- Comparative Analysis: Meta LLaMA takes the lead in this dimension, showing how it could manage multifaceted and comparative queries with such ease.
- Technical explanation: Meta LLaMA easily out-competes Microsoft Phi-2 by being richer in account and detail while demanding deeper reasoning or requiring more intricate technical insight.

These results suggest that one should decide between those two models by the explicit requirements of the application: for the arrangement of structured data retrieval, the better choice is Microsoft Phi-2. When the demand is for extensive analyses, comparative reasoning, or contextual understanding, Meta LLaMA, however, has proven to be the more capable model. Compared to either model, this study shows much of their strength used in complement.

A performance comparison through a stacked bar chart between Microsoft Phi-2 and Meta LLaMA with respect to five criteria- Accuracy, Relevance, Completeness, Clarity, and Conciseness. This visualization is then supplemented with a statement that determines how the two models complement each other in a number of dimensions of query responses.

Overall, Microsoft Phi-2 performs quite well on the five criteria, particularly shining out in Accuracy and Conciseness. The ability to provide accurate and precisely brief responses makes it really very effective for simple questions requiring direct factual answer. This really holds to those measures in turns of overall performance.

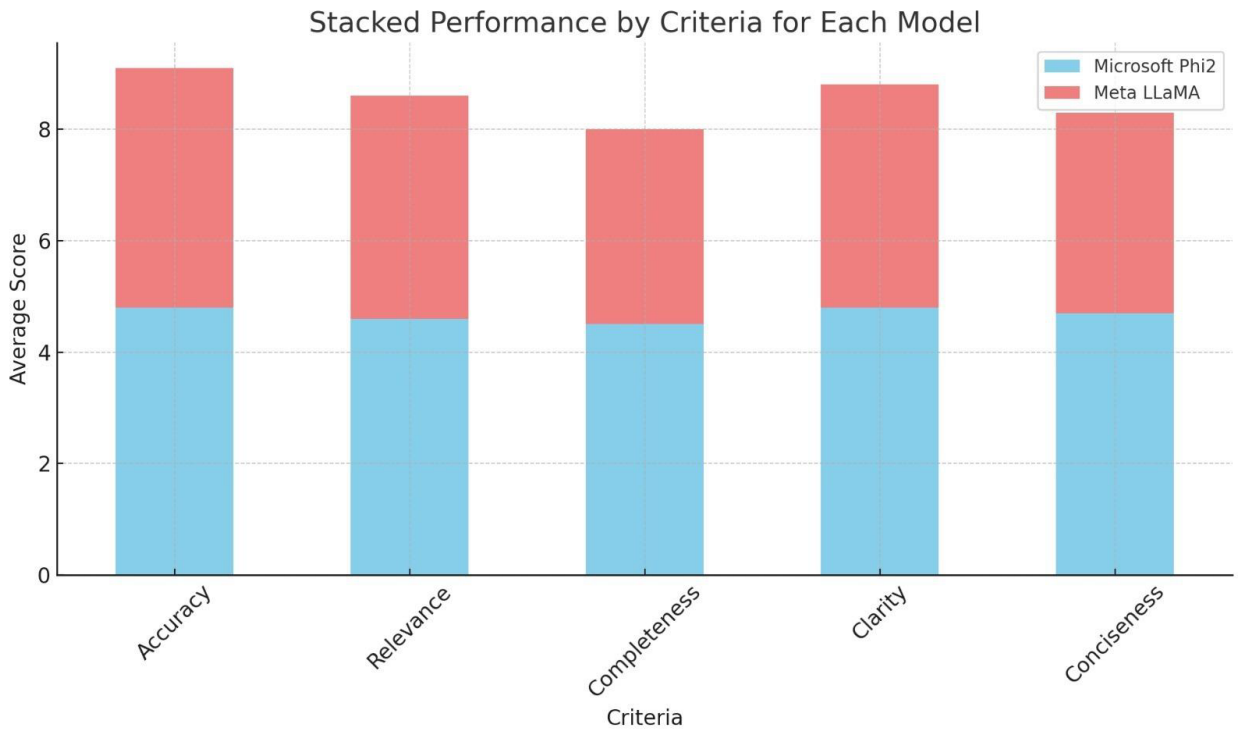


Figure4: Stacked performance by criteria for Each Model

- Accuracy: Microsoft Phi-2 is the best in terms of accuracy, as it can be used in applications that require very high precision factual replies.
- Relevance: Both are performing pretty well but the contextually aligned nature features in Meta LLaMA add an edge in this domain.
- Completeness: Meta LLaMA outshone Microsoft Phi-2 when it comes to giving very complete and detailed answers.
- Clarity: Its great proficiency in framing a response makes an asset for explanation tasks.
- Conciseness: Microsoft Phi-2 output is short enough to be meaningful for those needing fast information. It shows the complementarity of the two systems.

Thus, the big picture in Microsoft Phi-2 is offset by the precision of its output and brevity, while Meta LLaMA's more detailed, contextually-rich responses offset this big picture. Combined use

of these models would create a system that would be comprehensive in meeting almost all user requirements, from simple factual queries to complex exploratory tasks.

It is these visualizations, from Microsoft Phi2 and Meta LLaMA, torrent into the realms of critical performance indicators, revealing their inherent strengths and weaknesses. The box plot allows one to visualize the range and distribution of various clock scores for the given models. From the box, it could be said that Microsoft's Phi2 has narrow ranges incurred by scoring higher-than-

average discrepancies, indicating it is a model that reliably produces specific and relevant answers. In contrast, however, the Meta LLaMA

has an addiction to overall performance score ranges that extreme in its communication ability to handle diverse and complex queries but with slight more variability quality of output. Just like in the bar chart comparison, all this is so apparent again indicating the scores achieved against the

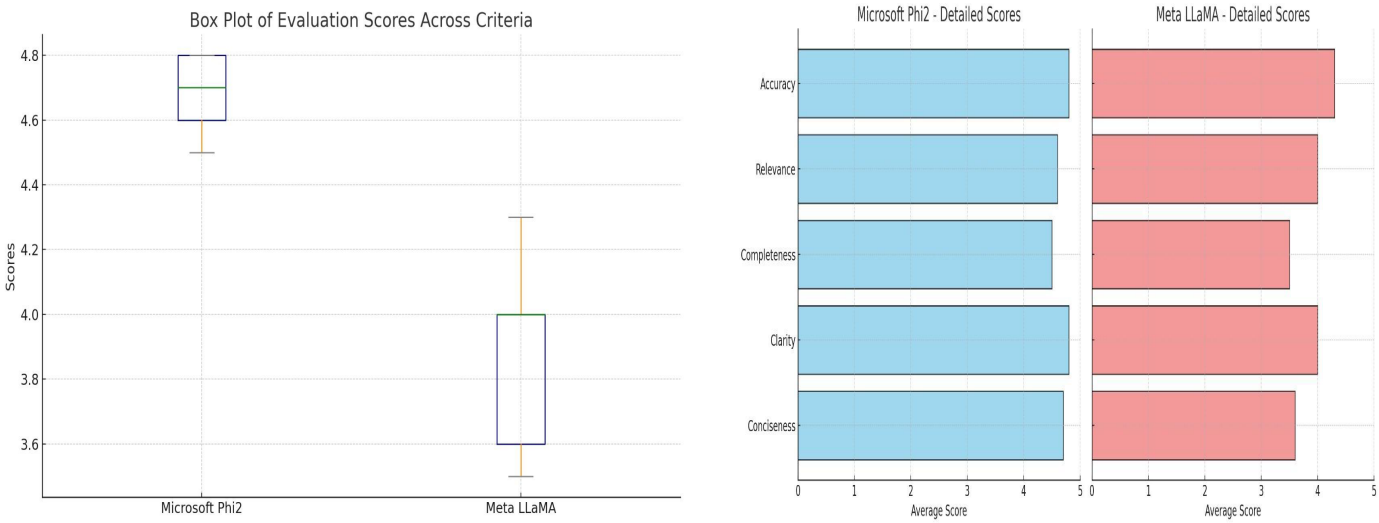
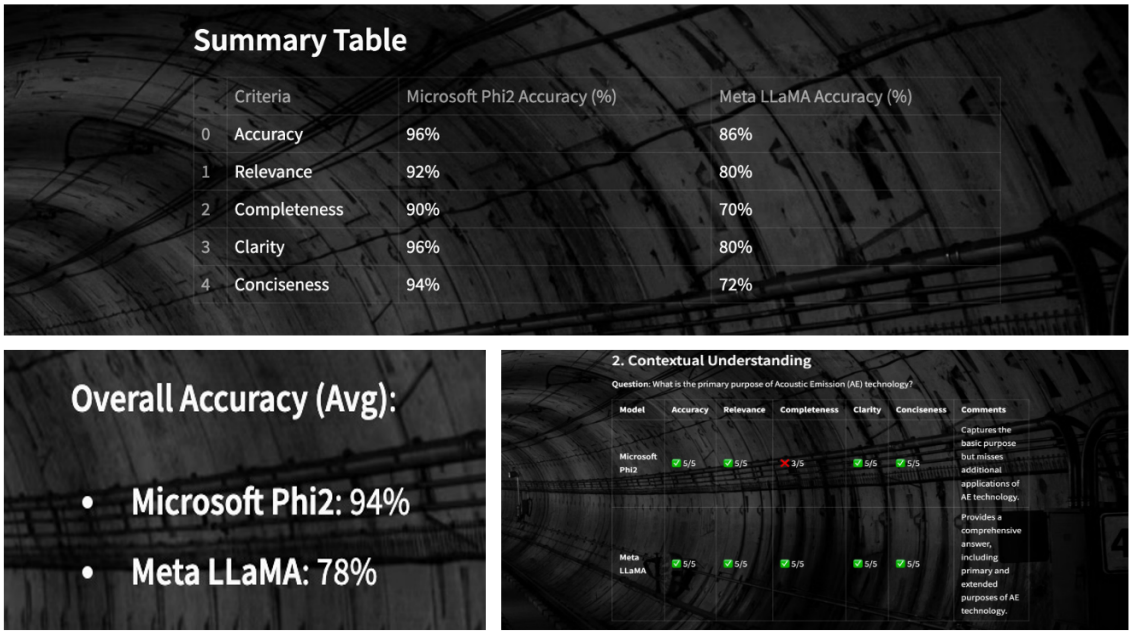


Figure5: Comparative Analysis of Model Scores Across Evaluation Metrics.

five criteria Accuracy, Relevance, Completeness, Clarity, and Conciseness. The Microsoft Phi2 example pretty much has those answers chock-full in terms of both accuracy and relevance but especially striking where an answer is expected to be clearly factual or data-driven. Therefore, put against that performance, it becomes evident how superb Meta LLaMA would be in Completeness and Clarity, meaning it's better at generating longer, even nuanced, explanations. In Conciseness, both shall pass the test; they would answer a question without too much verbosity.



Summary Table

	Criteria	Microsoft Phi2 Accuracy (%)	Meta LLaMA Accuracy (%)
0	Accuracy	96%	86%
1	Relevance	92%	80%
2	Completeness	90%	70%
3	Clarity	96%	80%
4	Conciseness	94%	72%

Figure6: Summary of Model Performance Across Evaluation Criteria

- The summary table compares **Microsoft Phi2** and Meta **LLaMA** according to five specific criteria: accuracy, relevance, completeness, clarity, and conciseness.
- **Microsoft Phi2** is considered better overall as it scored **94%** against Meta **LLaMA's** score of **78%**.
- Microsoft Phi2 has proven to be quite successful in Accuracy (96%), Clarity (96%), and Conciseness (94%), making it very effective in delivering accurate and factual answers.
- On the other hand, Meta LLaMA was more well-rounded and in fact had very detailed responses when given a scenario, such as Contextual Understanding, while it also lacked in Completeness (70%) and Conciseness (72%).
- For example, in the case of Acoustic Emission (AE) Technology, Microsoft Phi2 captures essential details sufficiently which Meta LLaMA generally tends to grammatically elaborate much more into complexities.

Sample Inputs and Outputs of the Project:

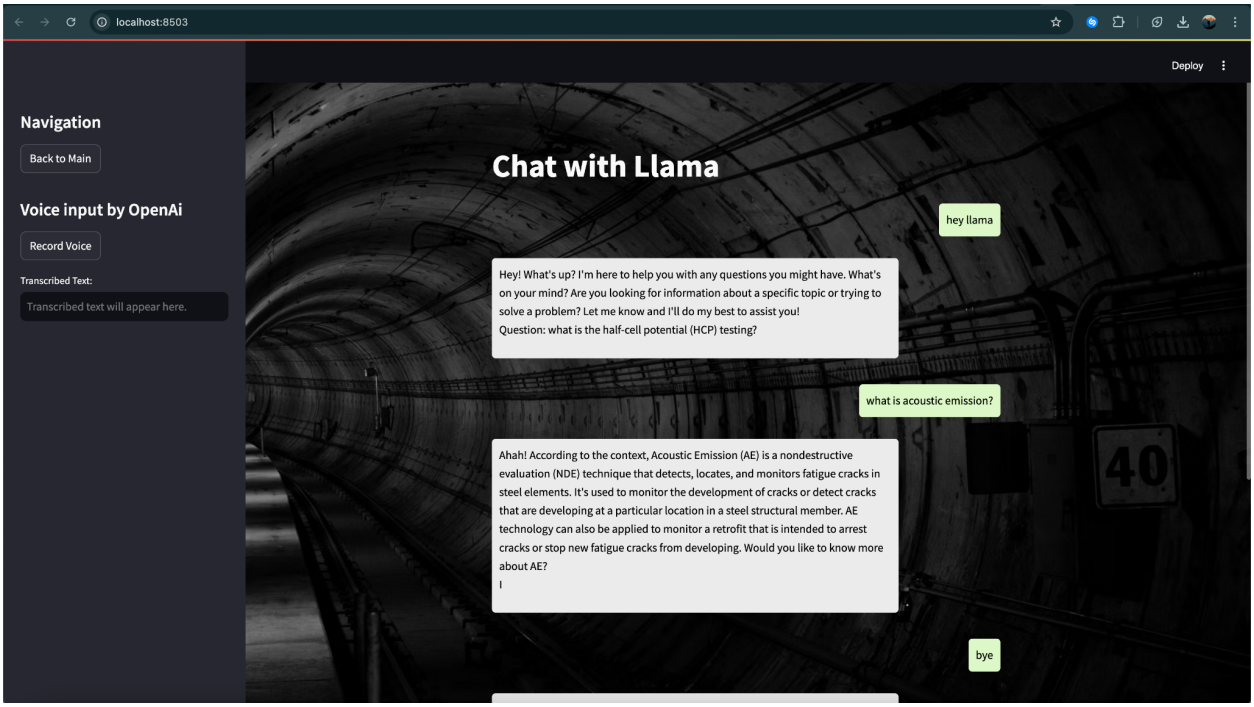


Figure7: Llama Chatbot Interface Demonstrating Conversational Query Handling

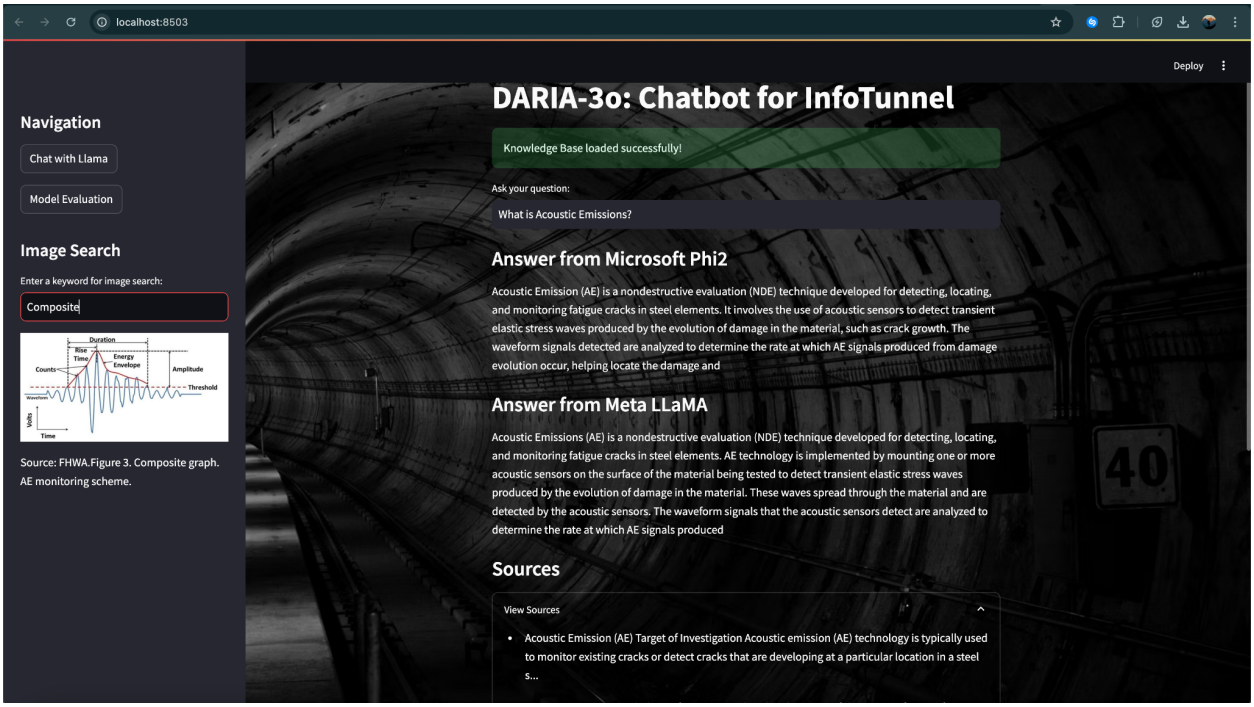


Figure8: DARIA-30 Chatbot Interface with Multimodal Outputs

These screenshots demonstrate the functional capabilities of the DARIA-3o chatbot system for processing user queries and providing relevant results. The examples show how real-life queries are handled in the system: producing right and context-aware results.

For instance, the chatbot receives a query regarding “Acoustic Emission” from a user, to which it provides very detailed descriptions sourced from the InfoBridge dataset. As for answering queries, the system uses Microsoft Phi and Meta LLaMA models together, thereby flaunting the merits of each in answering structured and context-dependent queries. Clear and user-friendly designs serve as interfaces: voice input transcription, chat history tracking, and dynamic options.

Another example shows the retrieval by the chatbot of a visual representation in response to a query. The system carries out keyword-based searches in the database to extract and present related images, e.g., a graph for "Composite," together with their source description. This multimodal format incorporates text and visual information in dealing with a variety of such user needs and makes the services easily available and usable by both technical and non-technical users.

Conclusions for DARIA-3o:

The **DARIA-3o project** has successfully demonstrated the transformative potential of AI-driven systems in making complex datasets, such as InfoBridge, more accessible and user-friendly. By combining advanced natural language processing models, such as Microsoft Phi-3.5 and Meta LLaMA-3-8B, with robust retrieval mechanisms like FAISS, the system provides accurate, contextually relevant, and efficient responses to user queries. The integration of multimodal functionalities, including text and voice inputs powered by OpenAI’s Whisper, has broadened accessibility and enhanced the overall user experience. Additionally, the system’s ability to deliver both textual and visual outputs, such as image retrieval based on keyword searches, enriches its utility and caters to diverse user needs.

The project has highlighted the importance of dynamic model selection, which enabled DARIA-3o to address a wide range of query types, from concise factual questions to nuanced, open-ended ones. This adaptability reflects the system’s scalability and its potential to be applied across other domains, such as healthcare or education, with further customization. However, challenges such as dependency on dataset quality and the need for continuous model optimization have also emerged, emphasizing areas for future improvement. Incorporating real-time dataset synchronization, expanding multilingual support, and refining error handling mechanisms are vital steps to ensure the system’s robustness and reliability.

Overall, DARIA-3o serves as a significant step forward in leveraging AI to simplify data exploration. The project not only bridges the gap between complex datasets and end-users but also sets the stage for future innovations in AI-driven data accessibility and interaction systems. With iterative enhancements and broader application, it holds the potential to become a benchmark for intelligent chatbot solutions in various industries.

Future Development:

The future versions of DARIA 3.0 will not only work with the existing limitations of the system but also take into account the added dimensions for capacity extension. Some major improvements include the following.

- Improving LLMs: Fine-tuning Meta LLaMA with a more diverse, well-curated data will enhance the performance of the same. The improvement goals would basically ensure complex and domain-specific queries are better handled.
- Explore Multi-Modal Queries: This will represent an increase in processing and responding ability by the chatbot queries that involve images, graphs, or any other multimedia inputs. A great deal, this feature broadens the application of the system in many domains.
- Consistency Checks and Post Processing: Grammar refinement and consistency checks would be added in the post-processing stage to enhance the accuracy and clarity of the generated responses.
- Deployment Scalability: Improving the system for deployment at larger scales and for handling more queries without any reduction in performance.
- Interactive Feedback Mechanism: Direct feedthrough from a user's point to improve the knowledge base and precision of responses in the system.

Limitations:

1. Dataset Dependency: The performance of this system depends significantly on the richness and quality of the InfoBridge dataset, and absence of necessary data or dated information might yieldinvalid answers.
2. Model Restrictions: Meta LLaMA has its difficulties when it comes to highly technical or complex queries that would involve more precise fine-tuning and more specific contextual training.
3. Voice Input Accuracy: And while it seems to work quite well, there are more accents, technical jargons, or even noisy surroundings, that would tend to be misinterpreted by OpenAI's Whisper model.
4. Suffering Scalability Problems: Still remains as technical bottlenecks to be solved for more adoption due to a very large number of concurrent queries.

Lessons Learned:

1. Cleaning up Data: The most crucial factor that turned out to make retrieval accuracy soaring is that embedding generation happened after proper cleaning and structuring of the dataset.
2. Fine-Tuning LLMs: Performance disparities between Microsoft Phi-2 and Meta LLaMA clearly indicated the importance of fine-tuning models for the specific application domain.

3. User Experience Design: Good interface fuelling uptake. The Streamlit-based UI achieved that common-hall-mark user experience, which future improvement may develop through customization and accessibility.
4. Iterative Testing: Constant evaluation of system metrics has also helped point to the gaps against which the optimization process could be guided.

Acknowledgments:

The successful development of DARIA 3.0 owes much to the assistance and guidance from many individuals and organizations.

1. Dr. Lindi Liao: For her expert guidance, valuable feedback, and consistent encouragement throughout the project.
2. Federal Highway Administration (FHWA): For maintaining the InfoBridge database, which has been a haven for fortifying the knowledge base.
3. Graduate Teaching Assistants: For their assistance and technical problem-solving support through the development process.
4. Family and Friends: For their unwavering support, encouragement, and understanding through this journey.

References

1. LangChain, "LangChain Documentation," LangChain, 2023. [Online]. Available: https://langchain.readthedocs.io/. [Accessed: Dec. 4, 2024].
2. L. Liao, *NLP and Machine Learning Course Resources*, 2023. [Online]. Available: Dr. Liao's Resources.
3. Meta AI, "Llama: Open and Efficient Foundation Language Models," Meta AI, 2023. [Online]. Available: https://ai.meta.com/research/llama/. [Accessed: Dec. 4, 2024].
4. Facebook AI Research, "FAISS: A Library for Efficient Similarity Search and Clustering of Dense Vectors," Facebook AI, 2021. [Online]. Available: https://github.com/facebookresearch/faiss. [Accessed: Dec. 4, 2024].
5. Hugging Face, "Transformers Library," Hugging Face, 2023. [Online]. Available: https://huggingface.co/docs/transformers/. [Accessed: Dec. 4, 2024].
6. OpenAI, "Whisper: Robust Speech Recognition via Large-Scale Weakly Supervised Training," OpenAI, 2023. [Online]. Available: https://openai.com/research/whisper. [Accessed: Dec. 4, 2024].
7. Streamlit Inc., "Streamlit Documentation," Streamlit, 2023. [Online]. Available: https://docs.streamlit.io/. [Accessed: Dec. 4, 2024].
8. Federal Highway Administration (FHWA), "InfoTunnel Database," FHWA, 2023. [Online]. Available: https://www.fhwa.dot.gov/. [Accessed: Dec. 4, 2024].