Group 41 - Proposal for CSE508 - Winter 2024

Arnav Goel, Aditya Pratap Singh, Ashutosh Gera, Medha Hira, Nalish Jain, Shikhar Sharma
{arnav21519,aditya20016,ashutosh21026,medha21265,nalish21543,shikhar20121}@iiitd.ac.in
IIIT Delhi, New Delhi
India

ACM Reference Format:
Arnav Goel, Aditya Pratap Singh, Ashutosh Gera, Medha Hira, Nalish Jain, Shikhar Sharma. 2024. Group 41 - Proposal for CSE508 - Winter 2024. In . ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In the current digital landscape, users face challenges in efficiently accessing relevant information and personalized assistance across different websites. The lack of a **unified assistant** leads to a fragmented user experience, impacting productivity and accessibility. To address this, our comprehensive browser extension acts as a co-pilot, seamlessly integrating features like a chatbot interface, multimodal content retrieval, versatile language support, semantic caching for performance optimization, and robust databases for chat and tab history storage.

2 PROBLEM IMPORTANCE

Addressing this problem is not just a convenience but a critical enhancement to the overall user experience in web interactions. The absence of a unified assistant across all website tabs poses a significant challenge, impacting user productivity and satisfaction.

Due to the lack of a similar tool issues faced are:

- **Fragmented User Experience**: The lack of a unified intelligent assistant leads to disjointed interactions and a suboptimal user experience. This hampers productivity, causing challenges in understanding, retrieving information, accessing personalized assistance.
- **Accessibility Concerns**: The absence of a cohesive solution hampers accessibility, making it challenging for users to find needed information, especially across multiple tabs, affecting their ability to fully leverage resources on various websites.
- **Complex Online Landscape**: In the era of vast digital information, navigating the increasingly complex online landscape has become daunting for users. The absence of a unified assistant exacerbates this challenge. Current hardcoded chatbots contribute to user frustration by providing only default, pre-fed answers.
- **Need for Personalized Assistance**: Users desire personalized support during online interactions. The lack of an intelligent assistant hinders their ability to receive tailored information and assistance, impacting the relevance and specificity of support.
- **Impact on Multilingual Interactions**: In the realm of online interactions involving diverse languages, a versatile language model is crucial. Absence of such capability introduces language barriers, constraining effective communication for users.

Our solution integrates a chatbot interface, an information retrieval system, a sophisticated language model, and key technologies to enhance user experience significantly. It addresses the limitations of current systems, especially those dependent on chatbots, by extending the scope of answers beyond the confines of a specific website's domain knowledge. This capability addresses a common user frustration: the reliance on guesswork or the expensive fallback of routing queries to human operators. Our browser extension eliminates these issues by offering quick, context-sensitive responses, streamlining the browsing experience without the usual delays or inaccuracies. Moreover, it reduces the need to escalate queries to human operators, thereby increasing operational efficiency and providing a more cost-effective solution. This strategy significantly diminishes the system's reliance on human intervention, making it more autonomous and efficient.

3 TECHNOLOGIES & ALGORITHMS TO BE USED

- **Frontend for the ChatBot Interface**: **ReactJS** will be used to create an interactive user interface that contains a seamless and easy-to-use UI, giving the users a quality experience.
- **Backend**: The set of scraped/retrieved webpages will be stored across multiple vector databases. Possible Vector DBs that we are considering using include Pinecone and Milvus. The URL to each individual vector DB will be stored as a foreign key in a parent Mongo database.
- **Retrieval Algorithms for Ranking and Matching**: Similarity search algorithms include an Approximate Nearest Neighbor search. Similarity matching will be done using Cosine Similarity and Jaccard similarity scores, taking an average of the two.
- **Ranking**: We will aim to investigate probabilistic and vector space models for ranking and use algorithms like PageRank. Ranking can also be done on the metadata which can be generated synthetically.
- **Audio Processing (Audio inputs and for audio retrieval)** Whisper (OpenAI) can be used for ASR. Bhashini API is for ASR and TTS for Indian Languages.
- **Creating a Chrome Extension**: This is the toolkit we will use for creating a Chrome extension. (Toolkit)
- **Framework for applications of LLM**: Langchain will be used to call LLM APIs and use them with our retrieval framework.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

4 NOVELTY

In comparison to the existing systems, our proposed idea offers a significantly enhanced user experience through a more robust context derived directly from the open tab. Our approach focuses on seamless integration and a stronger contextual understanding, ensuring better alignment with user preferences. A key feature is the personalization of user content, achieved through real-time data scraping on logged-in pages, ensuring that the information presented is both relevant and up-to-date. The system's capacity to tap into the entire webpage, as well as the specific section the user is viewing, offers a more nuanced understanding of the user's context. Additionally, our system supports multiple tab functionalities with a history that extends up to three previous tabs for enhanced pilot control.

Moreover, our idea stands out with the integration of modality in both input and output channels. The inclusion of multimodal retrieval further distinguishes our system, enabling us to answer user queries using a combination of various modes of information retrieval. Finally, the implementation of page-based question prompts ensures that users are presented with relevant inquiries tailored to each page, fostering a more interactive and user-centric browsing experience.

5 RELATED WORK

Current products exist like `Microsoft Copilot` and `Perplexity AI` . In the initial phase of our pipeline, data extraction plays a pivotal role in determining the accuracy of the subsequent output. Previous works, such as the one by the authors **[1]**, have employed NLTK's punkt tokenizer to proficiently split sentences and tokenize them into Unicode word tokens. Additionally, the Body Text Extraction (BTE) algorithm, rooted in HTML tag distribution, is utilized to enhance the extraction process. The paper introduces three ensembles, representing novel state-of-the-art extraction baselines. Moreover, the authors have demonstrated a thorough approach by amalgamating and refining various existing human-labeled web content extraction datasets. This meticulous compilation creates a comprehensive benchmark for evaluation, offering valuable insights that prove instrumental in guiding our project forward.

Complex search tasks require more than support for rudimentary fact-finding or re-finding. The recent emergence of generative artificial intelligence (AI) and the arrival of assistive agents, or copilots, based on this technology has the potential to offer further assistance to searchers, especially those engaged in complex tasks. The authors **[4]** discuss the challenges and opportunities for researching, developing and deploying search copilots and ends by concluding that careful search interface design is required to help people quickly understand copilot capabilities, to unify search and copilots to simplify the search experience and preserve flow.

In the landscape of Multimodal Large Language Models (MM-LLM), the paper **[5]** introduces NExT-GPT, an innovative system designed for versatile interactions across diverse modalities, including text, images, videos, and audio. Unlike prior models, NExT-GPT achieves this by connecting a Large Language Model (LLM) with multimodal adaptors and distinct diffusion decoders. Notably, the model leverages well-trained encoders and decoders, requiring minimal parameter tuning, making it cost-effective and poised for potential expansion into more modalities. The introduction of modality-switching instruction tuning (MosIT) and a dedicated dataset for MosIT enhances NExT-GPT's cross-modal semantic understanding and content generation capabilities. This research makes a significant contribution by presenting NExT-GPT as a pioneering MM-LLM system, bridging modalities and paving the way for more human-like AI agents.

Conversational Information Seeking (CIS) has emerged as a new paradigm for search engines, in contrast to the traditional query-SERP paradigm, it allows users to express their information need by directly conversing with the search engine. **[3]** introduces a comprehensive approach to CIS by developing a pipeline, a dataset, and a model to enhance how information is retrieved in conversational systems. Their proposed pipeline comprises six sub-tasks: **intent detection, keyphrase extraction, action prediction, query selection, passage selection, and response generation**, aiming to closely integrate the traditional search engine functionalities with conversational AI characteristics. The dataset, named WISE (Wizard of Search Engine), is created in a wizard-of-oz setup to simulate human-human CIS conversations, capturing a wide array of information needs and conversational interactions. Furthermore, the paper proposes a modular end-to-end neural architecture tailored to each sub-task, allowing for both joint and separate training and evaluation.

An important aspect of our work involves augmenting large language models with retrieved passages to improve domain-specific question answering. **[2]** mentions how generative models for question answering can benefit from passage retrieval. It retrieves support text passages from an external source of knowledge, such as Wikipedia. Then, a generative encoder-decoder model produces the answer, conditioned on the question and the retrieved passages. We use the findings from the paper to support our idea behind using retrieval to better augment LLMs on user queries about a webpage.

6 EVALUATION METRICS

Retrieval Metrics: Precision | Recall | Normalized Discounted Cumulative Gain (NDCG) | Mean Reciprocal Rank(MRR@10)

Text Generation Metrics:BLEU (Bilingual Evaluation Understudy) | NIST (Normalized Information Score) | METEOR (Metric for Evaluation of Translation with Explicit Ordering)

Evaluation of our system will be two-fold. One for evaluating the quality and relevance of the retrieved passages (i.e. Retrieval Metrics). These will measure the relevance of the retrieved passage and the ranking done by our algorithm. The second would involve evaluating the quality of text generated.

7 POTENTIAL CONTRIBUTIONS

Our contribution in this work will be developing a pipeline capable of chatting with the user and retrieving answers based on queries on pages they are surfing. This will be provided by an extension that exists as an overlay of the browser. We also aim to involve multimodal retrieval by allowing Q&A over images and audio data. This will be supported by multiple multimodal input and output channels according to the query. Our contributions will be vital in making open-source products which can be helpful for understanding focused context and improve grounding.

REFERENCES

[1] Janek Bevendorff, Sanket Gupta, Johannes Kiesel, and Benno Stein. 2023. An Empirical Comparison of Web Content Extraction Algorithms. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (<conf-loc>, <city>Taipei</city>, <country>Taiwan</country>, </conf-loc>) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 2594–2603. https://doi.org/10.1145/3539618.3591920

[2] Gautier Izacard and Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, Online, 874–880. https://doi.org/10.18653/v1/2021.eacl-main.74

[3] Pengjie Ren, Zhongkun Liu, Xiaomeng Song, Hongtao Tian, Zhumin Chen, Zhaochun Ren, and Maarten de Rijke. 2021. Wizard of Search Engine: Access to Information Through Conversations with Search Engines. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (<conf-loc>, <city>Virtual Event</city>, <country>Canada</country>, </conf-loc>) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 533–543. https://doi.org/10.1145/3404835.3462897

[4] Ryen W. White. 2024. Tasks, Copilots, and the Future of Search: A Keynote at SIGIR 2023. *SIGIR Forum* 57, 2, Article 4 (jan 2024), 8 pages. https://doi.org/10.1145/3642979.3642985

[5] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. NExT-GPT: Any-to-Any Multimodal LLM. arXiv:2309.05519 [cs.AI]