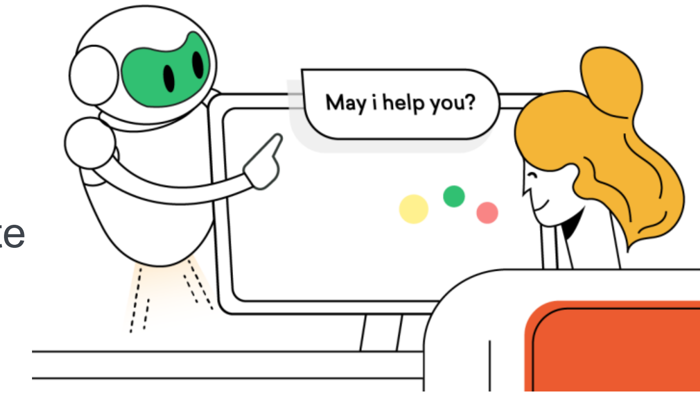
University of Stuttgart
Institute of Industrial Automation
and Software Engineering

Transforming Vehicle User Manuals into Interactive AI Chatbot Powered by Large Language Model

Speaker:Juntao Lin
Supervisor:Yuchen Xia
Data:09.09.2024

Contents

Introduction



- Motivation
  - machines understand and human language generate
  - Ability to handle a wide range of tasks
- Limitation
  - despite **generalization capabilities** models **lack specific domain knowledge**
- Objective
  - make general-purpose LLMs **specialized**
  - by transforming **vehicle user manuals into a knowledge base**
  - **RAG-based chat** assistant

Background

State of the art of LLM and RAG

Development Framework and Tools that support RAG

The dataset that support RAG

State of the art

LLM

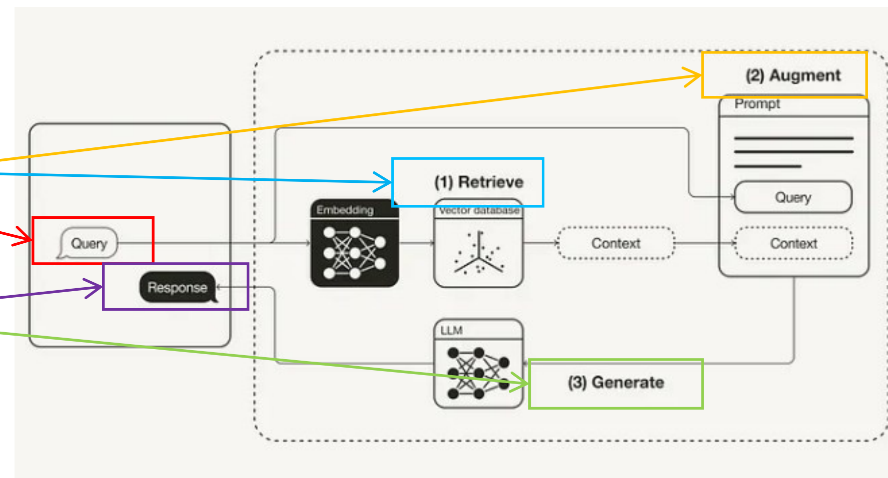
- Large Language Models (LLMs)
  - Overview
  - versatile tools in AI applications

Model Name	Publishing Agency	Parameters	Token Limit	Model Structure	Open Source
GPT-3.5	OpenAI	175B	4096	GPT-3	no
GPT-4	OpenAI	1.76T	128k	GPT-4	no
T5	Google	13B	-	T5-style	yes
PaLM	Google	540B	8192	GPT-style	no
Chinchilla	DeepMind	70B	-	GPT-style	no
LLaMA	Meta	7B-65B	2048-16k	GPT-style	yes

State of the art

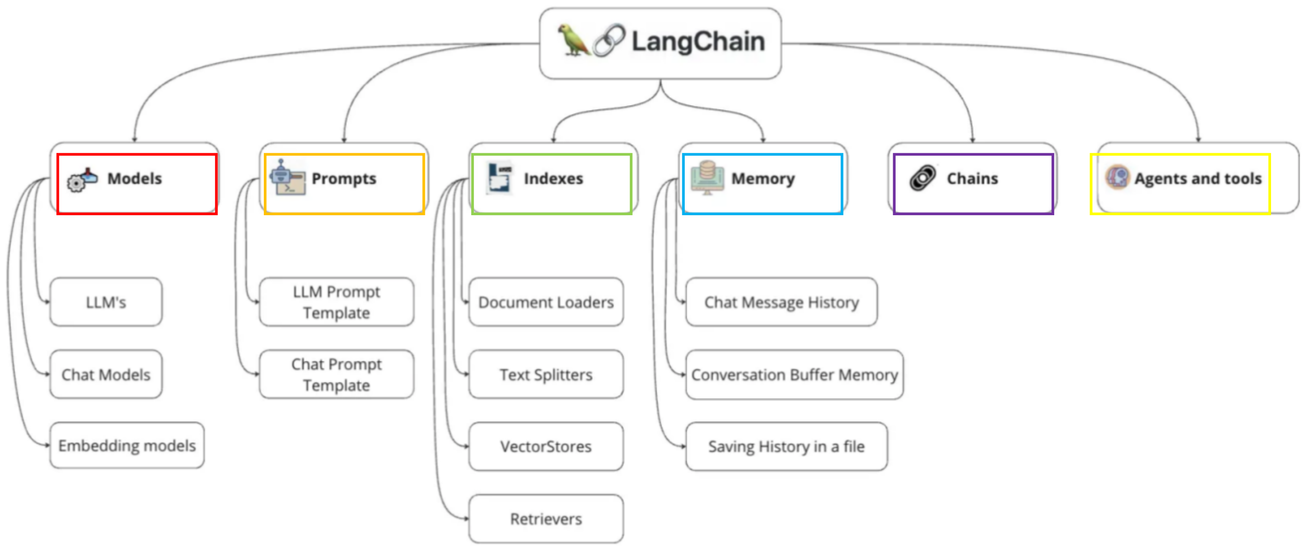
RAG

- Retrieval-Augmented Generation (RAG)
  - Addresse the limitations of LLMs in **domain-specific knowledge**
  - Allow to **retrieve relevant documents** from an **external database**
  - Provide **more precise** and **contextually relevant answers**
  - **Workflow** of the RAG pipeline.
    - Query
    - Retrieve
    - Augment
    - Generate
    - Response



Development Framework and Tools that support RAG

- Overview of development frameworks
  - **Langchain**: a **framework** designed for building applications powered by LLMs



  - **Langsmith**: **debugging, testing, and evaluating** the performance of these applications.
  - **Programming Language**:Python, HTML
  - **Development environment**:Pycharm

The dataset that support RAG

Vehicle User Manuals as a Knowledge Base

BYD ATTO 3 OM	2024/2/22 20:00	WPS PDF 文档	36,377 KB
BYD DOLPHIN OM	2024/2/22 19:59	WPS PDF 文档	31,632 KB
BYD HAN EV OM	2024/2/22 19:58	WPS PDF 文档	57,767 KB
BYD SEAL OM	2024/2/22 19:59	WPS PDF 文档	62,985 KB
MAZDA CX-5 OM	2024/3/10 21:33	WPS PDF 文档	17,789 KB
OPEL-AMPERA E OM	2024/2/22 20:06	WPS PDF 文档	6,857 KB
OPEL-MOKKA X OM	2024/2/22 20:09	WPS PDF 文档	7,023 KB
TANG EV OM	2024/2/22 19:57	WPS PDF 文档	63,090 KB
TESLA Model 3 OM	2024/2/22 20:25	WPS PDF 文档	8,365 KB
TESLA Model S OM	2024/2/22 20:24	WPS PDF 文档	8,570 KB
TESLA Model X OM	2024/2/22 20:26	WPS PDF 文档	10,452 KB
TESLA Model Y OM	2024/2/22 20:25	WPS PDF 文档	9,944 KB
VOLVO OM	2024/2/22 19:54	WPS PDF 文档	27,538 KB

- Comprehensive documents provided by manufacturers
  - operation
  - maintenance
  - safety features

Contents

Dataset preparation

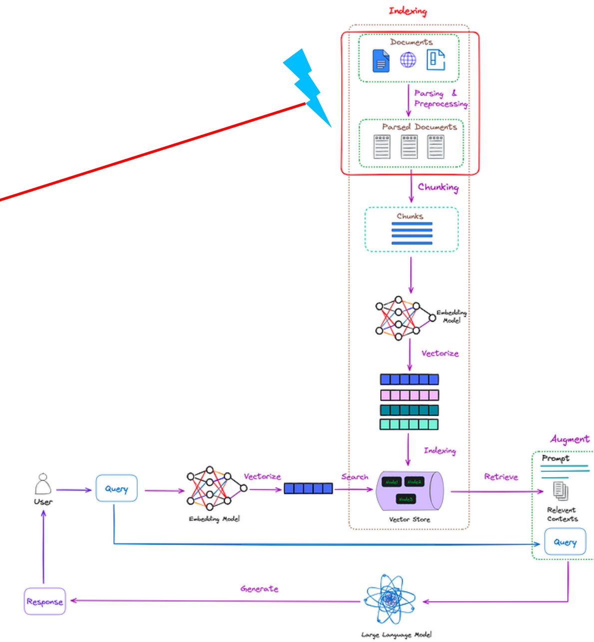
Chanllenges in PDF Parsing

Common Methods for Parsing PDF

Chunks of Raw Corpus

Chanllenges in PDF Parsing

- Inaccuracies in text extraction and layout recognition.
  - text misalignment
  - incorrect table parsing
  - loss of structural information during extraction

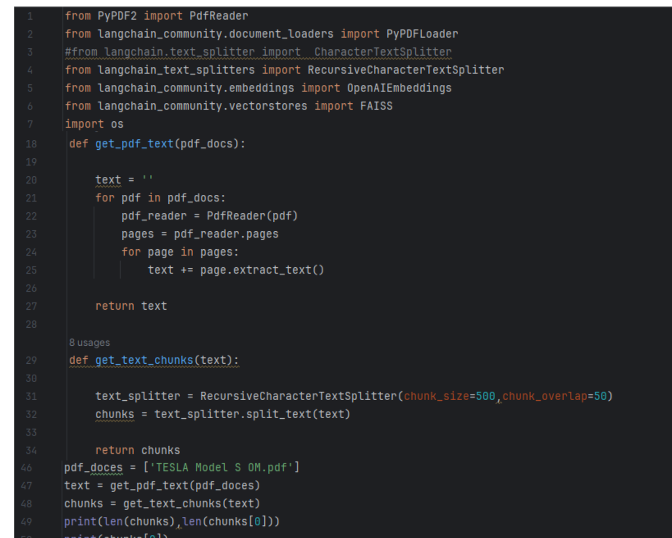


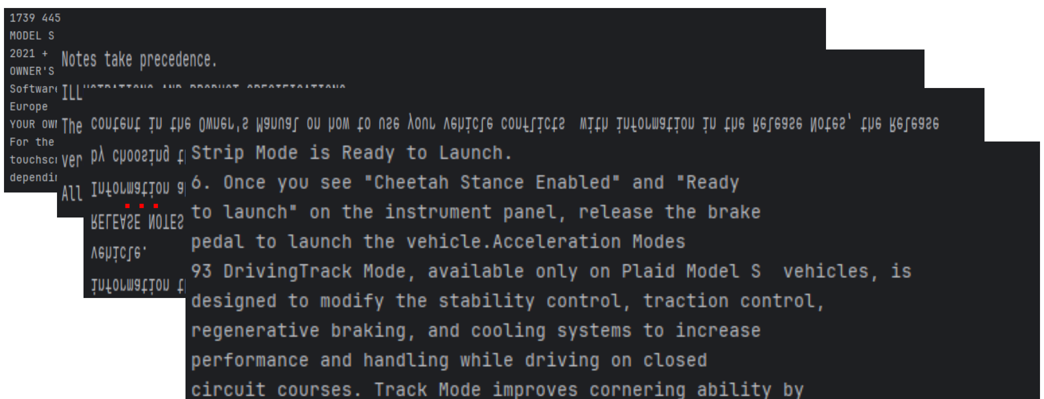
Common Methods for Parsing PDF

- Rule-based Approach:
  - rely on rules to extract content from PDFs.
    - for simple layouts
    - struggle with complex documents.
  - E.g.:
    - **pypdf,pdfplumber,ReportLab**
- Based on Deep Learning Models:
  - Leverage object detection and OCR models
    - better understand the structure of a document
    - require significant computational resources.
  - E.g.:
    - **Unstructured,Layout-parser,PP-StructureV2**
- Based on Multimodal Large Models:
  - combine text and image processing capabilities
    - provide a more comprehensive solution for parsing complex documents.
  - E.g.:
    - **GPT4-V,OCR model with GPT4/GPT3.5**

Chunks of Raw Corpus

- Import libraries:
  - PyPDF2,langchain_community.document_loaders,RecursiveCharacterTextSplitter
- Define functions:
  - ‘get_pdf_text’, ‘get_text_chunks’
- Process the PDF
  - print chunks





RAG Chatbot

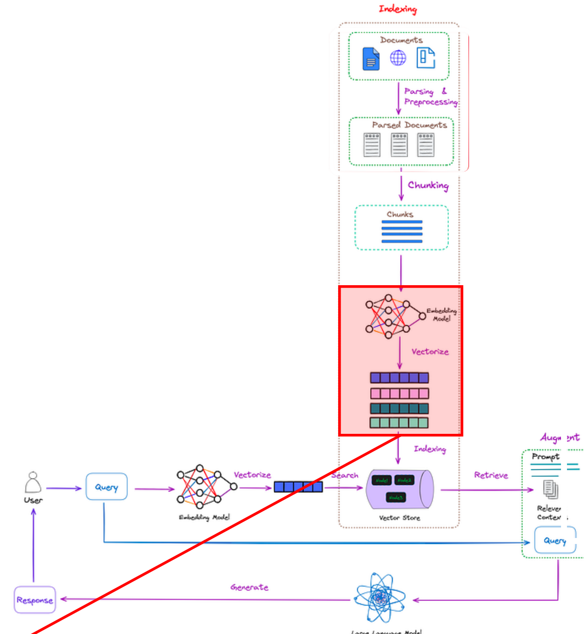
Embedding in Vector Database

Retriever Implementation

Prompt Engineering

Embedding in Vector Database

- Defining Function:
  - 'get_vectorstore'
    - Embeddings Model:text-embedding-3-small
    - Creating the Vector Store:FAISS
    - Return Statement



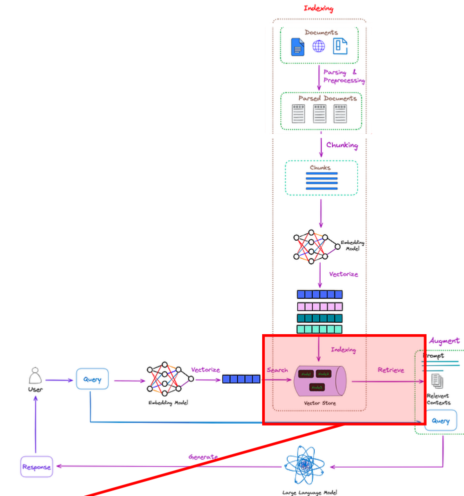

```
def get_vectorstore(chunks):

    embeddings_model = OpenAIEmbeddings(openai_api_key=openai_api_key)
    vectorstore = FAISS.from_texts(texts=chunks, embedding=embeddings_model)

    return vectorstore
vectorstore = get_vectorstore(get_text_chunks(get_pdf_text(['TESLA Model S OM.pdf'])))
```

Retriever Implementation

- Conversational Retrieval Chain
- Parameters
  - Chat_model:gpt3.5,llama2:7B,mistral:7B,llama3:8B
  - Search Type:Cosine Similarity
  - Search Parameters:return the **top 5** most similar results
  - Memory:remember previous interactions
  - Verbose:enables detailed logging or output

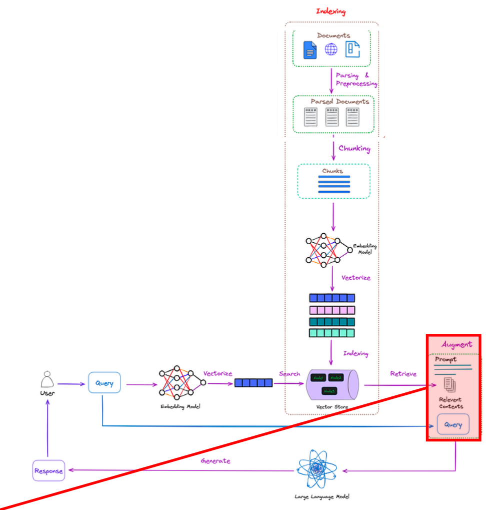



```python
conversation_chain = ConversationalRetrievalChain.from_llm(llm=chat_model,
                                                           retriever=vectorstore.as_retriever(search_type='similarity',search_kwargs={'k':5}),
                                                           memory=memory,
                                                           verbose=True)
```

Prompt Engineering

- How prompts are structured for a question-answering system related to car user manuals?
- System Prompt
  - Role Definition
  - Use of Context:
  - Handling Unknowns
  - Answer Length
  - Context Insertion
- User Prompt

```
system_prompt = f"""you are an assistant for question-answering tasks related to car user manuals.
                    use the following pieces of retrieved context to answer the user's question.
                    If you don't know the answer, just say 'I don't know'.
                    Keep your answer to 150 words or less.
                    \n\n
                    {context}"""
user_prompt = f"""user question is: {user_question}"""
```

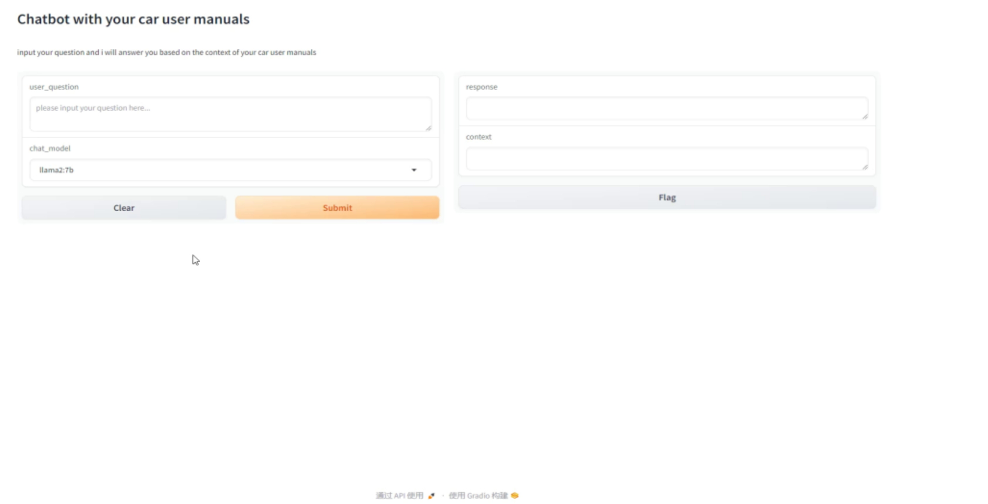


Achievements

User Application Interface of Chatbot

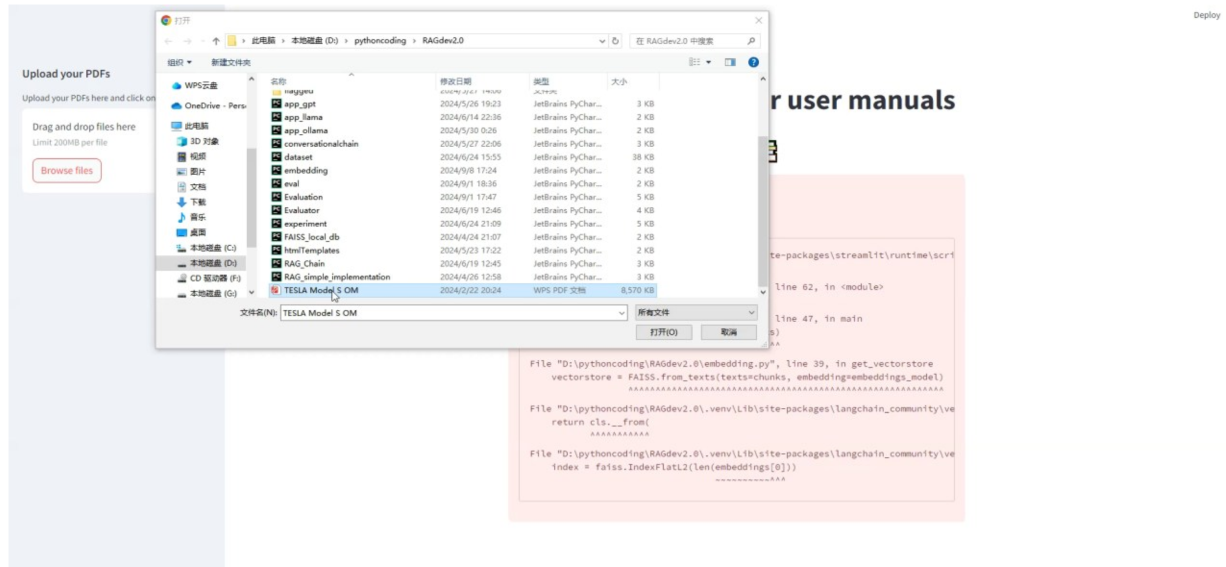
User Application Interface

- Tools Used:Streamlit and Gradio for interface development.
  - Chatbot developed using Gradio
    - User Input Section:
      - Text Input Box
      - Model Selection Dropdown
      - Submit and Clear Buttons
    - Response Section:
      - Response Box
      - Context Box



User Application Interface

- Tools Used:Streamlit and Gradio for interface development.
  - Chatbot developed using Streamlit
    - PDF Upload Section:
      - upload PDFs
      - File Management
    - Chat Interface:
      - Question Input
      - Response Display



Evaluation and Result Analysis

Evaluation Dataset preparation

Evaluation Indicators

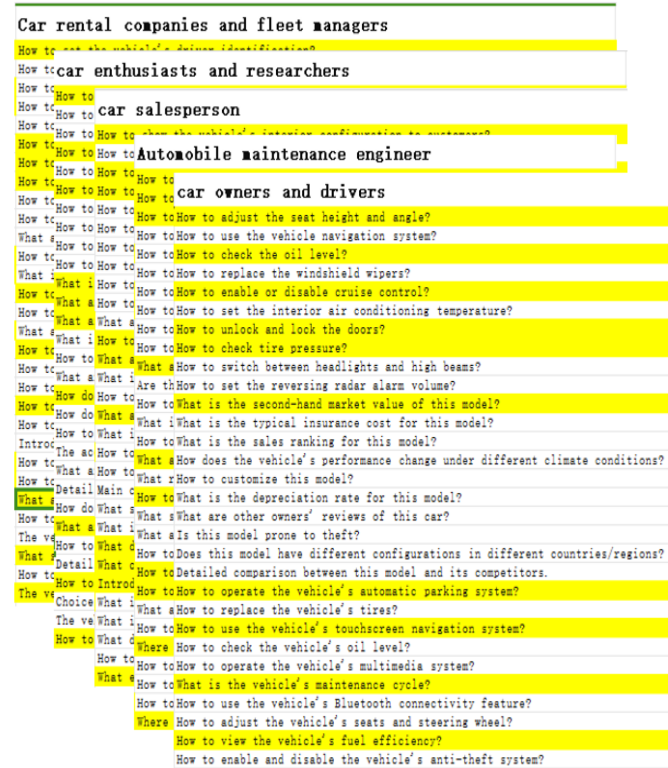
Overall Performance Scores of Chatbots

Performance Scores of Chatbots in 3 Querytypes

Low Score Attribution

Evaluation Dataset preparation

- Evaluation Dataset(**total 50 questions**)
  - five types of User Group
    - **Car owners and drivers** (10questions)
    - **Automobile maintenance engineers**(10questions)
    - **Car salespersons**(10questions)
    - **Car enthusiasts and researchers**(10questions)
    - **Car rental companies**(10questions)
  - three types of User Question
    - unanswerable questions(type1-15questions)
    - half-answerable questions(type2-17questions)
    - answerable questions(type3-18questions)



Evaluation Indicators

- Faithfulness **(Result correctness)**:
  - Measures the factual consistency of the generated answer given the context.
- Answer Relevancy**(Stick to the topic)**:
  - How relevant the answer is to the question.
- Context Relevancy**(Retrieve hit rate)**:
  - The relevance of the retrieved context to the original question
- Human evaluation: Overall evaluation based on answer helpfulness.

Overall Performance Scores of Chatbots

LLM integrated in chatbot	Result correctness (Faithfulness)	Stick to the topic (Answer_relevancy_score)	Human_evaluation_score
GPT3.5	0.67	0.70	0.56
Llama2:7B	0.57	0.87	0.67
Mistral:7B	0.73	0.70	0.58
Llama3:8B	0.47	0.40	0.69

- Complicated & controversial evaluation result
- Result correctness:mistral:7B scored the highest
- Stick to the topic: llama2:7B performed best
- Human evaluation:gpt3.5 were surprisingly not good

Performance Scores of Chatbots in 3 Querytypes

	Querytype	Result correctness	Stick to the topic	Retrieve hit rate
GPT3.5	unanswerable question	0.49	0.38	0.04
	half-answerable question	0.72	0.79	0.06
	answerable question	0.88	0.88	0.11
Llama2:7B	unanswerable question	0.34	0.74	0.04
	half-answerable question	0.64	0.90	0.06
	answerable question	0.71	0.95	0.12
Mistral:7B	unanswerable question	0.59	0.49	0.04
	half-answerable question	0.80	0.68	0.06
	answerable question	0.66	0.88	0.12
Llama3:8B	unanswerable question	0.24	0.12	0.04
	half-answerable question	0.54	0.56	0.06
	answerable question	0.49	0.53	0.12

Low score attribution

- Low score attribution and Manual labeling
  - Not retrieved(R1): the relevant information is not present
  - Missed the top ranked documents(R2):the relevant information exists but ranks too low
  - Not used by generative model(R3):correct information have not been used to produce responses
  - Noise in retrieved information(R4): irrelevant or low-quality content retrieved

Low Score Attribution

- Low Score Attribution For All half-answerable question

	R1: Not retrieved	R2: Missed the top ranked documents	R3:Not used by generative model	R4: Noise in retrieved information
Score 1: Result correctness	0.82	not observed	0.96	0.96
Score 2: Stick to the topic	not observed	1.00	not observed	0.88
Score3:Retrieve hit rate	0.70	0.94	0.91	0.46

- Low Score Attribution For All answerable question

	R1: Not retrieved	R2: Missed the top ranked documents	R3:Not used by generative model	R4:Noise in retrieved information
Score 1: Resualt correctness	0.61	not observed	0.75	0.93
Score 2: Stick to the topic	not observed	0.90	not observed	1.00
Score3:Retrieve hit rate	0.52	0.95	0.86	0.45

Low Score Attribution

- Comparison of low score attribution for half-answerable question of each model

LLM integrated in chatbot	scoreclass	Not retrieved	Missed the top ranked documents	Not used by generative model	Noise in retrieved information
GPT3.5	Resualt correctness	0.80	not observed	1.00	1.00
	Stick to the topic	not observed	1.00	not observed	1.00
	Retrieve hit rate	0.63	0.94	0.88	0.31
Llama2:7B	Resualt correctness	1.00	not observed	1.00	1.00
	Stick to the topic	not observed	1.00	not observed	0.00
	Retrieve hit rate	0.80	0.93	0.93	0.47
Mistral:7B	Resualt correctness	0.83	not observed	1.00	1.00
	Stick to the topic	not observed	1.00	not observed	1.00
	Retrieve hit rate	0.69	1.00	0.94	0.50
Llama3:8B	Resualt correctness	0.89	not observed	1.00	1.00
	Stick to the topic	not observed	1.00	not observed	0.86
	Retrieve hit rate	0.69	0.94	0.88	0.56

Low Score Attribution

- Comparison of low score attribution for answerable question of each model

LLM integrated in chatbot	scoreclass	Not retrieved	Missed the top ranked documents	Not used by generative model	Noise in retrieved information
GPT3.5	Resualt correctness	0.50	not observed	1.00	1.00
	Stick to the topic	not observed	1.00	not observed	0.88
	Retrieve hit rate	0.33	0.87	1.00	0.13
Llama2:7B	Resualt correctness	0.88	not observed	0.88	0.88
	Stick to the topic	not observed	0.00	not observed	0.00
	Retrieve hit rate	0.67	0.73	0.93	0.40
Mistral:7B	Resualt correctness	0.63	not observed	0.50	1.00
	Stick to the topic	not observed	1.00	not observed	1.00
	Retrieve hit rate	0.60	1.00	0.53	0.53
Llama3:8B	Resualt correctness	0.40	not observed	0.80	1.00
	Stick to the topic	not observed	0.88	not observed	1.00
	Retrieve hit rate	0.54	1.00	1.00	0.77

Summary and Outlook

Summary and Outlook

- Summary of Achievements:
  - successfully verified the **effectiveness of RAG technology**
  - transforming static vehicle manuals into **a dynamic, interactive LLM chatbot**
  - paves the way for more innovative applications of LLMs **in specialized domains**
- Challenges and Future Work:
  - focus on improving the chatbot’s ability.
    - enhance Contextual Understanding
    - dynamic Content Updating
    - feedback Loop Mechanism
    - improved Prompt Engineering

Question and Answer

University of Stuttgart
Institut of Industrial Automation
and Software Engineering

Thank you!

Juntao Lin

e-mail st176526@stud.uni-stuttgart.de
phone +49 (0) 711 685- 15773217446
fax +49 (0) 711 685-

University of Stuttgart
Institute of Industrial Automation and Software Engineering

Pfaffenwaldring 47, 70550 Stuttgart