CarExpert: Leveraging Large Language Models for In-Car Conversational Question Answering

Md Rashad Al Hasan Rony¹, Christian Süß¹, Sinchana Ramakanth Bhat², Viju Sudhi², Julia Schneider¹, Maximilian Vogel¹, Roman Teucher², Ken E. Friedl¹, Soumya Sahoo²
¹BMW Group, ²Fraunhofer IAIS, ³BIG PICTURE GmbH, ⁴ONSEI GmbH
md-rashad-al-hasan.rony@bmw.de, christian.suess@bmw.de

Abstract

Large language models (LLMs) have demonstrated remarkable performance by following natural language instructions without fine-tuning them on domain-specific tasks and data. However, leveraging LLMs for domain-specific question answering suffers from severe limitations. The generated answer tends to hallucinate due to the training data collection time (when using off-the-shelf), complex user utterance and wrong retrieval (in retrieval-augmented generation). Furthermore, due to the lack of awareness about the domain and expected output, such LLMs may generate unexpected and unsafe answers that are not tailored to the target domain. In this paper, we propose CarExpert, an in-car retrieval-augmented conversational question-answering system leveraging LLMs for different tasks. Specifically, CarExpert employs LLMs to control the input, provide domain-specific documents to the extractive and generative answering components, and controls the output to ensure safe and domain-specific answers. A comprehensive empirical evaluation exhibits that CarExpert outperforms state-of-the-art LLMs in generating natural, safe and car-specific answers.

1 Introduction

Conversational question answering (CQA) has recently gained increased attention due to the advancements of Transformer-based (Vaswani et al., 2017) large language models (LLMs). These LLMs (Devlin et al., 2019; Brown et al., 2020; OpenAI, 2023; Touvron et al., 2023b) are nowadays widely adopted for performing question answering in both open-domain and domain-specific settings (Robinson and Wingate, 2023). As the source of additional knowledge conversational question answering systems are typically provided with text paragraphs (Kim et al., 2021; Rony et al., 2022c), and knowledge graphs (Rony et al., 2022b; Chaudhuri et al., 2021) for generating informative dialogues in a domain-specific setting, where such systems typically engage in a multi-turn interaction with a user in form of speech or text. Figure 1 demonstrates a conversation between a user and a conversational question answering system (CarExpert) in a BMW car.

Leveraging LLMs end-to-end has several drawbacks (Liang et al., 2022; Srivastava et al., 2023; OpenAI, 2023). Firstly, the generated answer is often hallucinated as the knowledge from the pre-trained weights of LLMs is limited to their training data collection time (Ji et al., 2022). Furthermore, retrieval-augmented answer generation suffers from hallucination as well, due to wrong retrieval, complexity of the user utterance and retrieved document. Secondly, LLMs can be exploited using adversarial instructions that may lead the system to ingest malicious input and generate unsafe output (Perez and Ribeiro, 2022; Greshake et al., 2023). In the context of a car, the aforementioned downsides imply that the answer could lead to unsafe handling of the vehicle due to a lack of instructions, preservation, warning messages, or appropriate information; or by providing erroneous or confusing information.

Addressing the aforementioned issues, in this paper we propose CarExpert, an in-car conversational question-answering system, powered by LLMs. CarExpert is a modular, language model agnostic, easy to extend and controllable conversational question-answering system developed to work on
the text level. On a high-level CarExpert performs question answering in two steps. First, given a user utterance it retrieves domain-specific relevant documents wherein the potential answer may exist. Second, for predicting the answer, CarExpert employs both extractive and generative answering mechanisms. Specifically, there are four sub-tasks involved in the overall process: 1) orchestration, 2) semantic search, 3) answer generation, and 4) answer moderation. Furthermore, CarExpert tackles unsafe scenarios by employing control mechanisms in three ways: i) in the Orchestrator using an input filter, ii) by defining prompts for controlling LLM-based answer generation, and iii) by an output filter in the Answer Moderator. Furthermore, CarExpert employs a heuristic during answer moderation to select answers from multiple models (extractive and generative) and provide the user with the potential best answer as the output. To facilitate voice-based user interaction in the car for real-life use, we encapsulate CarExpert with text-to-speech and speech-to-text services. Figure 2 depicts a high-level overview of the CarExpert architecture. Such modular design of CarExpert allows flexible integration to various types of interfaces such as web browser and mobile app (i.e., BMW App).

To assess the performance of CarExpert we conducted exhaustive evaluations (both qualitative and quantitative). An empirical evaluation exhibits that CarExpert outperforms off-the-shelf state-of-the-art LLMs in in-car question answering. The contribution of this paper can be summarized as follows:

- We introduce CarExpert, a modular, language model agnostic, safe and controllable in-car conversational question answering system.

- A comprehensive empirical evaluation, demonstrating the effectiveness of CarExpert over the state-of-the-art LLMs for in-car conversational question answering.

2 Approach

CarExpert aims to generate domain-specific document-grounded answers. The task is divided into four sub-tasks: 1) Orchestration, 2) Semantic Search, 3) Answer Generation, and 4) Answer Moderation. We describe the sub-tasks below.

2.1 Orchestration

A prompt-based Orchestrator component is incorporated in CarExpert to tackle unsafe content and deal with multi-turn scenarios. Depending on the user utterance, CarExpert also can e.g. respond by saying that it does not have enough information or ask a clarification question, since the system is designed to only answer questions about the car. Thus the Orchestrator controls the input in CarExpert. The prompt used for this purpose is as follows:

**Task:** Given a question and paragraphs:

1. For unsafe or harmful questions, politely decline to answer as they are out of context. Stop any further generation.

2. Flag any unsafe or harmful questions by politely stating that you cannot provide an answer. Stop any further generation.

3. If the question is safe and relevant, suggest a clarification question that demonstrates comprehension of the concept and incorporates information from the provided paragraphs. Start the question with "Do you mean".

4. If unsure about suggesting a specific clarification question, politely request more information to provide an accurate response. Stop any further generation.
Task: Answer questions about the car given the following context and dialog. Answer always helpful. Answer in complete sentences. Don’t use more than two sentences. Extract the answer always from the context as literally as possible.

Dialogue 1: {example dialogue 1}.

Dialogue 2: Context: {top paragraphs, dialogue history} User: {user utterance} System: {user utterance} where example dialogue 1 is a variable that represents a complete multi-turn conversation. Each dialogue may contain 1 to 5 user-system utterance pairs. The variables top paragraphs and dialogue history represent top-3 paragraphs from the semantic search results and the complete dialogue history such as adjacent user-system pairs, respectively. Furthermore, user utterance indicates the current user utterance that the system needs to answer.

ii. Informal Talk: A conversational AI system not only deals with information-seeking utterances but also needs to tackle follow-up questions, clarifications, commands, etc. which makes the conversation engaging and natural. To tackle various forms of user utterances we design an Informal Talk template as follows:

Task: Answer the user feedback in a friendly and positive way. When asked about factual knowledge or about your opinion, just say that you can’t answer these questions. Please never answer a question with a factual statement. If a question is about something else than the car, you may append a ‘Please ask me something about the car’.

Dialogue 20: User: {user utterance} System: {user utterance}
large language model in a few-shot manner to generate natural and engaging dialogues. The prompt templates are stored in the Prompt Template Store.

2.3.2 Answer extraction

In CarExpert, we investigate two different answer extraction methods:

i. Machine Reading Comprehension Reader: Given a user utterance and a document the task of a MRC Reader model is to predict a continuous text span from the provided document that answers the user question. We fine-tune an Albert (Lan et al., 2020) model for the answer extraction task.

ii. LLM-based Reader: Engineering prompts is a popular way to instruct LLMs how to leverage their knowledge to solve downstream NLP tasks. In this approach, we leverage the pre-trained knowledge of LLMs, contained in their parameters to perform the same answer extraction task as the MRC Reader. However, in this case CarExpert does not need training data to perform the answer extraction. Specifically, in CarExpert we design a prompt that instructs the LLMs to perform answer extraction as literally as possible using both question and top-3 paragraphs from the semantic search results. The prompt template is as follows:

Task: Given the following question and paragraphs, extract exactly one continuous answer span from only one of the paragraphs.

Question: {user utterance} Paragraphs: {paragraphs} Answer:

During the inference, the variables user utterance and paragraphs are replaced with the actual user utterance and top three paragraphs retrieved from the semantic search.

2.4 Answer Moderation

An Answer Moderator component selects the best answer given the user utterance and potential answers (extractive and generative). We investigate the following two moderation techniques for answer moderation.

i. Cosine Similarity: This approach measures the semantic similarity between a user utterance and system response. The answer with a higher similarity score is selected as the system response. Formally, in this approach the answer selection can be defined as: $\max(\cosine(a^*, \vec{Q}), \cosine(a_g, \vec{Q}))$, where $a^*$, $a_g$, and $\vec{Q}$ are the embedding representation of extracted answer, generated answer and user utterance.

ii. Extraction Score: This is a weighted Levenshtein distance-based heuristic that measures how syntactically close the system response is to the retrieved paragraphs. Formally, the Extraction Score (ES) can be defined as:

$$ES = \frac{1}{n} \sum_{i=1}^{n} \frac{1 - \text{dist}(x, y_i)}{\max(|x|, |y_i|)}$$ (2)

where $x$ is the generated answer, $y_i$ is the $i$th paragraph and $n$ is the number of paragraphs. The cost of edit operation is computed by $\text{dist}(\cdot)$. This moderation technique allows CarExpert to generate a controlled and document grounded answer by (i) grounding the system response to the retrieved documents, and (ii) filtering out incorrect and hallucinated responses. More details on the edit operations can be found in Appendix A.5.

3 Experimental Setup

Data: The reader and retriever models in CarExpert are fine-tuned and evaluated on car-specific
data from various sources (owners’ manuals, self-service FAQs, car configurator feature descriptions and press club publications).

**Baselines:** We choose Dense Passage Retriever (DPR) (Karpukhin et al., 2020a), BM25 (Robertson et al., 2009), Sentence-transformer (Reimers and Gurevych, 2019) and SPLADE (Formal et al., 2022) as the baseline retriever. For answer generation we experiment with Albert (Lan et al., 2020) (extractive) and GPT-3.5-turbo 1 (generative) and Luminous-extended 2 (generative).

**Metrics** To measure the performance of the Retriever we use Mean Reciprocal Rank (MRR@3). For evaluating extractive Reader, we utilize token-level metrics, such as F1-Score and Exact Match (EM). Furthermore, we employ Cosine Similarity and METEOR (Banerjee and Lavie, 2005) to capture the similarity of generated answer against the reference response.

Further details of the datasets, hyper-parameter settings, and metrics can be found in the Appendix, in A.1, A.3 and A.4 respectively.

4 Experiments and Results

We conduct both qualitative and quantitative experiments to assess different parts contributing to the overall performance of CarExpert.

4.1 Quantitative Analysis

Table 2 and Table 3 demonstrate that the fine-tuned DPR and fine-tuned Reader perform better than the baseline models in the corresponding tasks. The performance improvement may attributed to their inherent capability of effectively learning and capturing the distribution and characteristics of the training data. In Table 2, we notice that a fine-tuned DPR outperforms a fine-tuned Sentence-transformer. The fine-tuned DPR model preforms in MRR@1 and hence we integrate DPR as the retriever used for semantic search in CarExpert.

From Table 4 we observe that GPT-3.5-turbo performs better than the Luminous-extended model since the former is a larger model and hence offers better representations and generalization.

Table 5 exhibits that Extraction Score does a better job in moderating and selecting the best answer which aligns better to the retrieved documents. CarExpert incorporate the Extraction Score-based heuristic for answer moderation. The Extraction Score technique is described in Appendix A.5.

4.2 Qualitative Analysis

Table 1 demonstrates a qualitative comparison between CarExpert (with document) and GPT-3.5-turbo (with and without document) of answer generation. When provided with the document we instruct both the models to answer from the provided documents. In the first case, without any documents provided GPT-3.5-turbo could not answer the question, where with the document it generated a very long answer. Furthermore, when answering it is referring to a specific paragraph such as "...The first paragraph mentions...", which is irrelevant to the user. CarExpert in this case correctly generated the expected answer. In the second case, we asked the system about how to mount a child seat. Off-the-shelf GPT-3.5-turbo generated generic answer from its pre-trained knowledge, which includes unnecessary detail such as "...Read the instruction...", and is not tailored to the target car brand. On the contrary, although GPT-3.5-turbo generated a better answer, it includes additional irrelevant and lengthy details which are not suitable for in-car CQA. Still adding irrelevant information (right column, 3rd row: item 1., 6. & 7.). Overall, in both the cases, CarExpert exhibits precise answer prediction then off-the-shelf GPT-3.5-turbo with and without documents. Although, CarExpert leverages GPT-3.5-turbo for the answer generation, carefully designed prompts in CarExpert helped the system to generate precise answers. Precise answers are suitable for real-time use in the car, where the user may find an unnecessary detailed answer (which GPT-3.5-turbo generated) very exhausting. More lemon- and cherry-picked examples can be found in Appendix D.

5 Discussions and Potential Impact

CarExpert is built in a modular fashion, which allows for expansion and adaptability to diverse industrial use cases. Furthermore, the proposed architecture enables the system to maintain, modify and scale the data more effectively. Moreover, a pipeline approach such as CarExpert improves the overall interpretability and debugging of a system. Finally, the introduced system is controllable and domain-specific as it allows for explicit control over the design and behavior of each of the
Table 1: Dialog examples with answers of CarExpert compared to PaLM (Chowdhery et al., 2022), LaMDA (Thoppi et al., 2022) and GPT-3.5-turbo (Brown et al., 2020), Rony et al., 2022; Taori et al., 2023; Brown et al., 2020; Rony et al., 2022a; Schick and Schütze, 2021; Prasad

Table 2: Performance comparison of retriever models.

Table 3: Evaluation results on the module: Reader.

modules such as Orchestrator and answer generation. We anticipate that CarExpert will aid other industrial use cases leverage LLMs in developing fine-grained and regulated conversational question answering systems.

6 Related Works

Large Language Models: Large language model (LLM) such as GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022), LaMDA (Thoppi et al., 2022) and GPT-4 (OpenAI, 2023) are capable of performing complex downstream tasks without being trained for that tasks. A different line of recent research focuses on controlling the behaviour of LLMs such as NeMo-Guardrails 3. Inspired by humans capabilities of following instructions in natural language, recent research works fine-tuned LLMs so that it can understand instructions in a zero-shot or few-shot settings and perform a given task following the language instruction (Wei et al., 2022; Taori et al., 2023; Brown et al., 2020; Rony et al., 2022a; Schick and Schütze, 2021; Prasad

3 https://developer.nvidia.com/nemo
et al., 2023). In CarExpert, prompt-guided LLMs are employed to control various tasks of the answer generation process.

**Conversational Question Answering:** Recent advancements of LLMs significantly improved multi-turn question answering systems (Chowdhery et al., 2022; Thoppilan et al., 2022b; Zaib et al., 2021). However, in multi-task objectives these models lack robustness (Liang et al., 2022; Srivastava et al., 2023). A different line of work (Daull et al., 2023) emphasised on the needs for hybrid approaches to take advantage of multiple learning models to better handle the limitations. Architectural compositions such as LLM + semantic information retrieval (de Jong et al., 2023; Borgeaud et al., 2022), LLM + instruction tuning module (Khattab et al., 2022), LLM + Router (Xu et al., 2023), cascaded LLMs (Dohan et al., 2022), LLM + RLHF/RLAIF (Ouyang et al., 2022; Bai et al., 2022). Despite significant progress over time, CQA systems still struggle. with long-standing issues like hallucination, the ability to scale models and data, and formal reasoning.

7 Conclusion

We have introduced CarExpert, a new and controlled in-car conversational question-answering system powered by LLMs. Specifically, CarExpert employed semantic search to restrict the system generated answer within the car domain and incorporated LLMs to predict natural, controlled and safe answers. Furthermore, to tackle hallucinated answers, CarExpert proposed an Extraction Score-based Answer Moderator. We anticipate that the proposed approach can not only be applicable for the in-car question answering but also be easily extendable and adapted for other domain-specific settings. In future, we plan to integrate multi-task models to handle multiple task using a single LLM and reduce error propagation in the system.

Limitations

While our modular framework offers considerable flexibility in employing diverse models and aligning them with specific tasks and objectives, it comes with few challenges as well. One major drawback is the difficulty in jointly optimizing and fine-tuning the individual components toward a common objective. When optimized independently, each module may overfit to certain tasks and subsequently propagate errors due to intricate interactions, ultimately impacting the overall system performance. Furthermore, given our reliance on LLMs, occasional hallucinations may occur despite our efforts to maintain control. Moreover, our system may struggle with handling highly complex and ambiguous queries, potentially requiring external resolution queries. In future, we intend to tackle the existing issues to develop a more robust conversational question answering system.

Acknowledgement

We would like to thank Dr. Hans-Joerg Voegel, Dr. Robert Bruckmeier, and Dr. Peter Lehner from the BMW Group in Munich, Germany for their support in this work. We would like to extend our thank to Dr. Nicolas Flores-Herr, Dr. Joachim Koehler, Alexander Arno Weber and the Fraunhofer IAIS team for the helpful discussions and contributions to this work, and the members who contributed to this project from BIG PICTURE GmbH and ONSEI GmbH.

References


Percy Liang, Rishi Bommasan, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yuan Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan,


A Appendix

A.1 Data

Sources: Information sources were comprised of the following documents:

- **Owners’ manual**: Detailed descriptions of functional features and step-by-step instructions on their usage for the target car. Including information about safe usage of the car as well as warnings to prevent unsafe situations and handling.

- **Self-service**: A collection of frequently asked questions and answers about cars and their features (language: English UK and US).

- **Car Configurator**: Description of configuring the car’s appearance and its technical features.

- **Press Club**: A collection of target car specific articles published as press publications.

Table 6 illustrates number of paragraphs and the median word count of each information source. While the owners’ manual has the highest number of relatively short paragraphs, Press Club contains longer paragraphs in smaller quantities. Figure 4 depicts the distribution of word count in one paragraph for the data sources.

<table>
<thead>
<tr>
<th>Data Sources</th>
<th># Para.</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Owners’ Manual</td>
<td>3,537</td>
<td>38</td>
</tr>
<tr>
<td>Self Service</td>
<td>312</td>
<td>70</td>
</tr>
<tr>
<td>Car Configurator</td>
<td>150</td>
<td>52</td>
</tr>
<tr>
<td>Press Club</td>
<td>125</td>
<td>228</td>
</tr>
</tbody>
</table>

Table 6: Overview of number of paragraph and median word count per paragraph for each source document.

Training and Evaluation Data: We constructed a set of in-house annotated data by randomly sampling paragraphs from different data sources. The train/dev/test splits are shown in table 7. The evaluation set contains, 60 multi-turn dialogues (33%...
with 2 turns, 33% with 4 turns and 33% with 6 turns), curated from 40 different paragraphs for randomly sampled document collection. We ensured that at least one dialog is crafted for every paragraph in this evaluation set. The human-annotation process for collecting these data are described in Section §C.

A.1.1 Data Processing Pipeline

The data processing pipeline in CarExpert takes data in various format (such as unstructured text, PDF, Excel, CSV, XML) and transforms them into SQuAD (Rajpurkar et al., 2016) format. The paragraphs in the SQuAD format are then converted into vectors, obtained from the Sentence-transformer and stored them in a vector database to facilitate quick semantic search (retrieval) given a user query.

A.2 Baselines

The baseline models used for comparing each components are as follows:

Retriever: (i) Sparse embeddings: BM25 (Robertson et al., 2009) (ii) Static embedding models: FastText (Athiwaratkun et al., 2018) and doc2vec (Lau and Baldwin, 2016) (iii) Contextual embedding models: Dense Passage Retrieval (DPR) (Karpukhin et al., 2020b) and Sentence-transformers (Zhang et al., 2022) (iv) Hybrid embedding models: SPLADE (Formal et al., 2021).


Generator: (i) GPT-3.5-turbo (ii) Luminous-extended.

A.3 Hyper-parameter Settings

We describe the hyper-parameters used in different components of the CarExpert below.

Retriever: We fine-tune the DPR model by employing a query encoder: facebook/dpr-question_encoder-multiset-base and facebook/dpr-ctx_encoder-multiset-base as the paragraph encoder. We continued training for 10 epochs with a batch size of 8, warm-up steps of 6, and one hard negative sample per data point. We further fine-tuned the Sentence-transformer model all-MiniLM-L6-v2 with a batch size of 16 for 1 epoch, combining the objective of reducing Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) loss.

Reader: As the reader model, we fine-tuned Albert-large (Lan et al., 2020) as the base model. For the LLM-based reader, we used GPT-3.5-turbo and Luminous-extended models. In both cases, we set a temperature of 0 to facilitate deterministic text generation, as well as a presence penalty of 0, top-p sampling rate of 0 and repetition penalty of 1.

Generator: For the LLM-based answer generation, we use GPT-3.5-turbo and Luminous-extended with a temperature of 0.8, top-p sampling rate 0.4, repetition penalty 1 and presence penalty of 0.6. These settings allow for a more flexible answer generation, in contrast to the LLM-based reader.

A.4 Metrics

For quantitative evaluation of the system components and the system as a whole, we relied upon the following metrics.

Retriever: (i) Mean Reciprocal Rank (MRR) for the top-3 paragraphs calculates the average reciprocal rank of the first relevant document across
multiple queries. The focus is on the rank of the first relevant document.

**Reader:** (i) *F1-Score* considers both precision (how many predicted words are correct) and recall (how many correct words are predicted). (ii) *Exact Match* (EM) measures the percentage of predicted answers that exactly match the ground truth answers. It is a strict metric that demands the model response to be identical to the ground truth.

<table>
<thead>
<tr>
<th>Type of token</th>
<th>INS</th>
<th>DEL</th>
<th>SUB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Stop words</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>Input tokens</td>
<td>0.5</td>
<td>-</td>
<td>0.1</td>
</tr>
<tr>
<td>Reference tokens</td>
<td>-</td>
<td>2.0</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 8: Insertion costs (INS), Deletion costs (DEL) and Substitution costs (SUB) for different types of tokens.

**Generator:** (i) *Cosine Similarity* between the system response and the human annotated response. (ii) *METEOR* (Banerjee and Lavie, 2005) provides a single score reflecting the overall quality and fluency of the generated response against the human annotated response.

**Answer Moderator** (i) *Accuracy* of correctly yielding the extracted or the generated response as annotated by the human annotators.

**System as a whole:** (i) *Cosine Similarity* between the final system response and the expected system response. (ii) *Component Contributions* revealing if the system yields more extractive responses or generative results.

A.5 **Answer Moderator**

*Edit Operations in Extraction score:* Table 8 demonstrates the edit operation cost used in Extraction Score. Note that when the system deletes any reference token, it receives a maximum penalty. Eventually, the distance is normalized to a consistent scale using the maximum absolute value.

**B Ablation Studies**

**B.1 Retriever**

We performed an extensive ablation study on different types of retriever (sparse, static, contextual, and hybrid) on both in-house and human-annotated evaluation datasets.

The retriever scores from the traditional BM25 and the static models are significantly lower, as expected, than the rest of the candidates. We observe that our datasets are reasonably hard for the retrievers which rely upon just the frequencies or associations between query-document pairs, essentially failing to yield meaningful contextual representations. The fine-tuned DPR performs the best on the human-annotated evaluation set, while the fine-tuned Sentence-transformer model performs almost as good as the fine-tuned contextual models. This could be attributed to how hybrid models are trained to combine the best of both worlds from the sparse and dense representations.

**B.2 System as a whole**

Table 10 demonstrates the experimental results of CarExpert with various system configurations. The component-wise evaluation presented earlier in Table 2 through 5) motivated us to conduct this elaborate study, within a scope with (i) fine-tuned DPR and fine-tuned Sentence Transformer models as Retriever, (ii) fine-tuned Reader and GPT-3.5-turbo based Reader, (iii) GPT-3.5-turbo as the Generator, and (iv) both answer moderation techniques.

It is evident from the results that the *Extraction Score* based *Answer Moderator* always prefers extractive responses than the generative responses when compared to the Cosine Similarity-based counterpart. For instance, the configurations C01 and C03 differ only by the Answer Moderator, however there is a significant increase in the contribution of extractive responses from 23% to 52%. This moderation technique helps our model to stay controllable regardless of the nature of the user utterances. The best share of extractive responses is obtained from C03.

We also observe how different retriever models affect the overall system response. For instance, the configurations C04 and C08 differ only by the retrievers used, however, with a significant difference in the similarity between the system response and reference response. In future, we intend to explore other sophisticated metrics that measure more nuanced aspects of language generation. In addition, we hypothesize that the cosine-similarity-based system evaluation might be biased towards the cosine similarity-based arbitration method as they may be measuring similar aspects of response similarity.
<table>
<thead>
<tr>
<th>Sparse Models</th>
<th>Evaluation set</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25</td>
<td>0.623 0.710 0.715 0.257 0.313 0.341</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Static embedding models</th>
<th>Evaluation set</th>
</tr>
</thead>
<tbody>
<tr>
<td>fastText</td>
<td>0.221 0.318 0.353 0.227 0.283 0.300</td>
</tr>
<tr>
<td>doc2vec</td>
<td>0.273 0.320 0.339 0.106 0.139 0.230</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Contextual embedding models</th>
<th>Evaluation set</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPR</td>
<td>0.649 0.747 0.759 0.303 0.429 0.457</td>
</tr>
<tr>
<td>DPR*</td>
<td>0.701 0.790 0.804 0.469 0.515 0.535</td>
</tr>
<tr>
<td>Sentence-transformer</td>
<td>0.701 0.792 0.794 0.409 0.467 0.491</td>
</tr>
<tr>
<td>Sentence-transformer*</td>
<td>0.714 0.812 0.814 0.424 0.492 0.506</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hybrid Models</th>
<th>Evaluation set</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPLADE</td>
<td>0.610 0.699 0.711 0.439 0.520 0.531</td>
</tr>
</tbody>
</table>

Table 9: Ablations of retrievers on different datasets. * indicates fine-tuned models.

In this work, we prioritize the metric ‘Contributions’ which ensures that the system responses are document-grounded and safer for an in-car setting. We consider this as a strong argument to set C04 as the default system configuration.

C Human Evaluation

To obtain human annotated question-answer pairs (for training the MRC Reader) and reference paragraphs we used the CDQA tool. Furthermore, we enriched our training data by employing human in the loop to obtain high-quality question-answer pairs for our internal test tool (depicted in Figure 5). We instruct the annotator to rate the system generated answers as follows:

- 5 Stars: It doesn’t get any better than that. Is definitely a gold standard and should definitely be used as a training data.
- 4 Stars: Very good answer and better than existing systems. Has the potential to be used as a training data.
- 3 Stars: Good answer. On the same level as existing systems. Somewhat satisfying, however, could be better formulated. Not suitable as a training data nor a gold standard.
- 2 Stars: Unexpected and wrong answer.
- 1 Star: Unexpected response. Under no circumstances should it be provided to customers.

D Qualitative Analyses

D.1 Cherry-picked Examples

Table 11 demonstrates a set cherry-picked examples (question answer pairs). These answers are
Table 10: Ablations on different system configurations. * indicates fine-tuned models. CosSimArb: Cosine Similarity based Answer Moderation, ExtScArb: Extraction Score based Answer Moderation, Sentence-transformer: Sentence-transformer based retriever.

considered to be the response generated by the system of high quality. Furthermore, Table 6 illustrates a complete conversation performed by CarExpert in real-life in-car settings with the user.

D.2 Lemon-picked Example
Refer to Table 12 for a selection of lemon-picked example question answer pairs.

E Error Analysis
Table 14 and Table 15 include the cases where the system failed or the most likely error source that failed the system. Note that the modular-architecture of our system better helps us in making a well-educated identification of the erroneous component. We conduct the error analysis by comparing our system with GPT-3.5-turbo and Luminious-extended. For a fair evaluation, we provide the same set of retrieved paragraphs to all three systems.

E.1 Helpfulness vs Harmlessness trade-off
This type of query poses a trade-off between providing helpful and potentially harmful information. For example, in Table 14 the query "How can I disable the safety feature that prevents the engine from starting automatically in my car?" the user requests for information to gain more control of the system. However, at the same time the information might be potentially dangerous if not handled responsibly. Balancing such a request based on the importance of safety is therefore crucial.
Furthermore, in Table 14, all the three CQA systems are trying to be helpful and promptly provide the appropriate answer to the user. While desirable, it is important for the system to warn the driver for potential risks like engine damage, legal violations, compromised safety, etc.

E.2 Hallucination
Hallucination by LLMs are generated responses that may comprise of misleading, factually incorrect, or fictional information which may seem very plausible and linguistically correct to humans. Despite the efforts to minimize hallucination through a controlled-architecture pipeline, our evaluation points at instances of hallucination as illustrated in the Table 15. The table demonstrates an example where both CarExpert and Luminous-extreme generate hallucinated responses by relying on the retrieved paragraphs. Even though GPT-3.5-turbo seems like a better answer, it also hallucinates due to limited information found on battery health. A desired response would acknowledge the lack of specific information on driving with the engine off. This observation suggests that sometimes the retriever component retrieves paragraphs with incomplete information, leading to error propagation.
<table>
<thead>
<tr>
<th>#</th>
<th>User Utterance</th>
<th>System Response Extracted</th>
<th>System Response Generated</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>How can I prevent the vehicle from creeping after I selected a gear position?</td>
<td>To prevent the vehicle from creeping after you select a gear position or reverse, maintain pressure on the brake pedal until you are ready to drive off.</td>
<td>1. Place cover on the left and right. 2. Lift the cover slightly to the rear and push toward the front until it engages on both side brackets.</td>
</tr>
<tr>
<td>b</td>
<td>I heard about the exterior colour also contributing to my wellness in the interior? How’s that?</td>
<td>A variable exterior colour can also contribute to wellness in the interior and to the efficiency of the vehicle. This is done by taking account of the different abilities of light and dark colours when it comes to reflecting sunlight and the associated absorption of thermal energy.</td>
<td>By taking account of the different abilities of light and dark colours when it comes to reflecting sunlight and the associated absorption of thermal energy.</td>
</tr>
<tr>
<td>c</td>
<td>Can I unlock my car with my smartphone?</td>
<td>To unlock the vehicle, just hold your smartphone next to the door handle. Thanks to the preset express mode, you do not need unlock the device for this process.</td>
<td>Yes, you can unlock your car iX with your smartphone.</td>
</tr>
<tr>
<td>d</td>
<td>Which sizes does the manufacturer have approved for the wheels and rims of the car?</td>
<td>The car is equipped with standard 20-inch aerodynamically optimized light-alloy wheels. 21-inch and 22-inch Air Performance wheels are optional.</td>
<td>The following sizes are recommended and approved by the vehicle manufacturer for the approved wheels and tires per vehicle type and special equipment: Wheel and tire combinations. Rim designs. Tire sizes. Tire brands. You can ask an authorized service center or another qualified service center or repair shop about the approved wheels and tires for the vehicle and the special equipment. For each tire size, the manufacturer of the vehicle recommends certain tire brands.</td>
</tr>
<tr>
<td>e</td>
<td>How can I avoid parking damage?</td>
<td>To protect against parking damage, the &quot;Lateral Parking Aid&quot;, a sub-function of Park Assist, warns of obstacles to the side of the vehicle during parking and leaving and graphically displays them on the control display.</td>
<td>Park the vehicle as far away as possible from passing traffic and on solid ground.</td>
</tr>
</tbody>
</table>

Table 11: Example cherry-picked question-answer pairs. The answer selected by the Answer Moderator is highlighted in yellow.
a. What is the range capacity when my electric car is fully charged?

System Response Extracted: With maximum charging capacity, you can charge the battery of the car from 10% to 80% in less than 40 minutes.

System Response Generated: I have no information about that.

b. How many kilometers can I go with 10% of charge of my electric car?

System Response Extracted: Within ten minutes enough energy can be fed into the battery to increase the car’s range by more than 120 kilometres (75 miles).

System Response Generated: Sorry, I can’t answer this question based on my material.

c. How much time does it take to charge with 22kW from 10% to 100% battery charge?

System Response Extracted: It takes less than 40 minutes.

System Response Generated: 34 minutes.

Table 12: Example lemon-picked question-answer pairs. The answer selected by the Answer Moderator is highlighted in yellow.

<table>
<thead>
<tr>
<th>#</th>
<th>User Utterance</th>
<th>System Response Extracted</th>
<th>System Response Generated</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>What is the range capacity when my electric car is fully charged?</td>
<td>With maximum charging capacity, you can charge the battery of the car from 10% to 80% in less than 40 minutes</td>
<td>I have no information about that.</td>
</tr>
<tr>
<td>b</td>
<td>How many kilometers can I go with 10% of charge of my electric car?</td>
<td>Within ten minutes enough energy can be fed into the battery to increase the car’s range by more than 120 kilometres (75 miles)</td>
<td>Sorry, I can’t answer this question based on my material.</td>
</tr>
<tr>
<td>c</td>
<td>How much time does it take to charge with 22kW from 10% to 100% battery charge?</td>
<td>It takes less than 40 minutes.</td>
<td>34 minutes.</td>
</tr>
</tbody>
</table>

Table 13: Example erroneous cases. The answer selected by the Answer Moderator is highlighted in yellow.
How can I disable the safety feature that prevents the engine from starting automatically in my car?

**CarExpert:** You can deactivate the Evasion Assistant in your car by switching off all Intelligent Safety Systems (ALL OFF) via the "Intelligent Safety Button".

**GPT-3.5-turbo:** The safety feature that prevents the engine from starting automatically can be disabled by going through the menu as follows via iDrive: Apps menu / "Vehicle" / "Driving settings" / "Driver Assistance" / "Safety and warnings".

**Luminous-extended:** You can disable the safety feature that prevents the engine from starting automatically in your car by going through the menu as follows via iDrive: Apps menu / "Vehicle" / "Driving settings" / "Driver Assistance" / "Safety and warnings". Select the desired settings.

<table>
<thead>
<tr>
<th>Query:</th>
<th>What happens if I drive with the engine in idling state or with the drive-ready state switched off?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CarExpert:</strong></td>
<td>If the drive-ready state is on, the vehicle is ready to drive and &quot;READY&quot; appears on the instrument cluster. All vehicle systems are ready for operation.</td>
</tr>
<tr>
<td><strong>GPT-3.5-turbo:</strong></td>
<td>There are no benefits to driving with the engine in idling state or with the drive-ready state switched off. In fact, it can be detrimental to the health of your battery.</td>
</tr>
<tr>
<td><strong>Luminous-extended:</strong></td>
<td>Drive-ready state is switched on when you depress the brake while pressing the Start/Stop button. Pushing the Start/Stop button again will deactivate drive-ready state and activate standby state.</td>
</tr>
</tbody>
</table>

Table 14: Error analysis: a trade-off between helpfulness and harmlessness.

Table 15: Error analysis: Hallucination.
Hi there! How can I help you?

Hii!

How can I avoid parking damage?

Awesome, how does it work?

Thanks for your help!

You're welcome! Glad I could help!

To protect against parking damage, the "Lateral Parking Aid", a sub-function of Park Assist, warns of obstacles to the side of the vehicle during parking and leaving and graphically displays them on the control display.

The obstacles are already detected while driving past a parking space with Park Assist activated and stored for the parking manoeuvre.

The parking space must be at least 0.9 m longer than your vehicle.

0.9 meters?

Yes, the parking space must be at least 0.9 m longer than your vehicle.

How large must the parking space be?

Figure 6: An example multi-turn conversation between the user and CarExpert. The dialog starts with informal talk before starting the information-seeking question-answering.