From Recall to Creation: Generating Follow-Up Questions Using Bloom's Taxonomy and Grice's Maxims

Archana Yadav[†*], **Harshvivek Kashid**[†*], **Pushpak Bhattacharyya**[†]
Medchalimi Sruthi[◇], B JayaPrakash[◇], Chintalapalli Raja Kullayappa[◇], Mandala Jagadeesh Reddy[◇]
[†]Indian Institute of Technology Bombay, India
[◇]Hyundai Motor India Engineering, India
(archanaqre@gmail.com, harshvivek@cse.iitb.ac.in, pb@cse.iitb.ac.in)

Abstract

In-car AI assistants enhance driving by enabling hands-free interactions, yet they often struggle with multi-turn conversations and fail to handle cognitively complex follow-up questions. This limits their effectiveness in real-world deployment. To address this limitation, we propose a framework that leverages Bloom's Taxonomy to systematically generate follow-up questions with increasing cognitive complexity and a Gricean-inspired evaluation framework to assess their Logical Consistency, Informativeness, Relevance, and Clarity. We introduce a dataset comprising 750 human-annotated seed questions and 3750 follow-up questions, with human evaluation confirming that 96.68% of the generated questions adhere to the intended Bloom's Taxonomy levels. Our approach, validated through both LLM-based and human assessments, also identifies the specific cognitive complexity level at which in-car AI assistants begin to falter information that can help developers measure and optimize key cognitive aspects of conversational performance.

1 Introduction

Large language models (LLMs) have transformed chatbots, enabling more natural and responsive interactions than rule-based ones. They are now common in customer service, education, tutoring, and entertainment, where they retrieve information and generate content through conversational interfaces. Despite these advances, many commercial AI assistants still struggle to answer user queries because of limited domain knowledge or cognitive constraints. This often leads to generic replies like "Sorry, I don't know," misinterpretations, or hallucinated facts, which frustrate users and reduce engagement, especially when questions demand more than simple recall.

Testing chatbots in the wild with manually crafted questions does not scale. It cannot support rapid iterations across Volume (large question sets), Variability (diverse domains), or Velocity (fast turnaround). Relying on an aggregate statistic—simply whether the chatbot can answer a question—overestimates performance and obscures where and why it fails (Ribeiro et al., 2020). Bloom's Taxonomy is a proven rubric for assessing cognitive skills. By issuing scaffolded questions at each level, we can systematically evaluate a chatbot's reasoning and application across increasing cognitive demands (see Figure 1).

Modern vehicles are increasingly integrated with LLMs to facilitate interactions between the in-car AI assistant and the driver. However, LLMs are not inherently designed for domain-specific tasks and lack automotive-specific knowledge and real-time data access, leading to generic failures. Our work focuses on evaluating LLM-powered in-car AI assistants. This is a high-stakes setting where misunderstandings or failures can impact safety and usability. By probing the assistant with our cognitively scaffolded methodology, we reveal its cognitive limitations and demonstrate that our evaluation approach generalizes to other LLM-powered chatbot applications.

Extensive work on question generation—ranging from template and statistical methods (Heilman and Smith, 2010), neural Seq2Seq models (Du et al., 2017) and semantic-graph approaches (Pan et al., 2020) to form-type balancing (Ghanem et al., 2022) ("how" vs "what") —focuses on generating high-quality questions rather than probing an LLM's cognitive abilities.

Prior studies have mapped benchmarks to Bloom's levels (Huber and Niklaus, 2025) and introduced Bloom-aligned tasks (Zoumpoulidi et al., 2024; Sun et al., 2024), but these rely on static, isolated questions or domain-specific

[*]Equal contribution

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 1322–1338

prompting. No existing work systematically probes LLMs with a sequence of single-turn follow-up questions that each increase in cognitive complexity. This leaves the model's stepwise reasoning across the complete taxonomy unexplored. Our approach fills this gap by assessing responses to cognitively scaffolded prompts, revealing weaknesses beyond surface-level accuracy.

In the in-car voice assistant domain, available datasets, such as KVRET (Eric and Manning, 2017), offer multi-turn dialogues but do not include follow-up questions that escalate in cognitive complexity. No corpus is explicitly designed to evaluate how an in-car AI assistant navigates successively harder prompts along Bloom's hierarchy.

Traditional evaluation metrics such as BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2019) depend on surface-level text similarity to ground-truth sentences, and thus fail to assess the nuanced quality of follow-up questions as experienced in real conversations. They cannot measure whether a question is truly relevant to the driver's task, whether it conveys new information, or whether it is phrased clearly and truthfully. Moreover, reference-based evaluation demands expensive human annotations or gold-standard follow-ups, which limits scalability across diverse driving scenarios. (RQUGE (Ge et al., 2023), an example of Gricean Maxims' implementation for evaluating questions, which evaluates only the previous turn).

To overcome these shortcomings, we turn to **Grice's Maxims**—the conversational principles of **Quantity**, **Quality**, **Relation** and **Manner**—as a natural rubric for evaluating follow-up questions in an in-car dialogue. We map each maxim to a reference-free metric:

- **Relevance** (Relation): Does the question focus on information pertinent to the driving context?
- **Informativeness** (Quantity): Does it introduce an appropriate amount of new, useful content?
- **Truthfulness** (Quality): Does the question logically follow from the previous context?
- **Clarity** (Manner): Is it unambiguous and easy to understand?

These Grice-inspired, reference-free metrics are scalable, adaptable, and cost-effective for evaluating large question sets in diverse driving scenarios.

As a developer of an in-car AI assistant technology, it is crucial to identify where the assistant fails, understand its cognitive limitations, and determine the types of questions it struggles to answer. To address this, we propose a technique that leverages LLMs to generate follow-up questions based on Bloom's Taxonomy. By systematically increasing the cognitive complexity of these questions, developers can assess the assistant's reasoning capabilities and pinpoint its limitations. Crucially, we avoid multi-turn dialogues where each follow-up depends on the assistant's previous answer. Chaining questions in this way can conflate errors, as a flawed response early on can derail the reasoning path and obscure the model's actual capabilities. Instead, we design each follow-up as a single-turn prompt, grounded only in the original context. This isolates the effect of increasing cognitive demand alone, avoids error propagation, and ensures that each question cleanly tests a distinct cognitive skill.

Our key contributions are:

1. **B-FQG Technique**: A Bloom's Taxonomy-based Follow-up Question Generation (FQG) method that produces follow-up questions by progressively increasing cognitive complexity—from recall to creation—without relying on previous responses from the in-car AI assistant powered by LLMs (Section 2.3).

2. **GriceWise**: A Grice's Maxims-inspired evaluation framework for follow-up questions. This reference-free method assesses questions based on logical consistency, informativeness, relevance, and clarity in multi-turn dialogues (Section 2.2).

3. **Blooms-FQ Dataset**: A human-annotated dataset comprising 750 seed questions and 3750 follow-up questions. Human evaluation confirms that 96.68% of the generated questions align with the intended Bloom's Taxonomy levels[1].

[1]Dataset link: https://huggingface.co/datasets/harshvivek14/Blooms-Followup-Questions

Figure 1: Illustration of Bloom's Taxonomy-based follow-up question generation for an in-car AI assistant. A Level 1 seed question is used to generate five follow-up questions that progressively increase in cognitive complexity, from basic recall to higher-order creative inquiries. The in-car AI assistant successfully answers simpler questions, but for certain higher-level queries, it defaults to a generic response such as "Sorry, I didn't get that!" highlighting its cognitive limitations. Responses marked with ✓ are correct or relevant, while ✗ indicate missing or evasive answers.

2 Methodology

In this section, we present our approach for generating follow-up questions that progressively increase cognitive complexity, guided by Bloom's Taxonomy. Our method, B-FQG (Bloom's Taxonomy-based Follow-up Question Generation), leverages both few-shot and zero-shot prompting to direct LLMs in producing follow-up questions that challenge in-car AI assistants at various cognitive levels. This systematic approach allows us to evaluate the cognitive capabilities of these systems and identify their limitations.

2.1 Seed Question Annotation

We construct the Bloom-FQ Dataset with 750 seed questions corresponding to Level 1 of Bloom's Taxonomy (Remember/Recall) for in-car AI assistants by converting a comprehensive list of supported commands—spanning Phone Calls, Sending Messages, POI Search, Media, Weather, Date and Time, Radio, Navigation Control, Climate Control, NLU Commands, and Automatic Temperature Control—into factual, minimal-reasoning questions (e.g., "Call John Smith" → "How do I make a call to John Smith?"). To ensure the dataset was non-redundant, we compared each pair of questions using semantic similarity and retained only one question from any pair with a similarity score above 0.95. This filtering process resulted in 750 unique seed questions (see Table 1 for domain-wise distribution). A second annotator then verified that each question adhered to Level 1 criteria—i.e., "what," "which," or "how" queries with a single, unambiguous answer, achieving 100% adherance to Level 1 of Bloom's Taxonomy. These verified seed questions serve as the foundation for our higher-level follow-up question generation.

Domain	# Qs	Domain	# Qs
Media	146	Climate Control	100
Phone	64	General Settings	63
POI Search	54	Navigation Control	50
Car Controls	47	Weather	41
Date and Time	41	NLU Commands	40
Car Manual	35	Sports	33
Radio	28	Messaging	8

Table 1: Domain-wise distribution of the 750 Seed Questions corresponding to Level 1 of Bloom's Taxonomy (Recall)

2.2 GriceWise: Gricean-inspired Evaluation Framework

We evaluate the follow-up questions using Grice's Maxims (Appendix A.1) to ensure capturing *Logical Consistency*, *Informativeness*, *Relevance* and *Clarity*. This ensures we are evaluating the questions beyond surface-level similarity.

2.2.1 Contextually-Relevant Gricean Scores

We define Q_1 as the seed question and $\{Q_2, Q_3, \ldots, Q_6\}$ as the sequence of follow-up questions. The context for the i-th follow-up question, denoted as C_i, includes all previous questions from Q_1 to Q_{i-1}, i.e., $C_i = \{Q_1, Q_2, \ldots, Q_{i-1}\}$

Logical Consistency (Maxim of Quality): To capture whether a follow-up question logically follows from the prior conversation, we adopt a *Natural Language Inference (NLI)* approach. Let C_i represent the prior context (including all preceding questions and answers), and let Q_i be the current follow-up question. We define the logical consistency score as the probability of the *entailment* label assigned by `roberta-large-mnli`[2]:

$$\text{LC}(Q_i \mid C_i) = \text{Entail}_{\texttt{roberta}}(Q_i, C_i)$$

A higher entailment score indicates that Q_i does not contradict or deviate from C_i, suggesting strong logical consistency. Conversely, a lower score implies that Q_i introduces inconsistencies or does not follow from the established conversation. This ensures that each follow-up question remains faithful to the context of the dialogue.

Informativeness (Maxim of Quantity): To capture the Informativeness of a question, we compute the conditional entropy of each follow-up question given the context of the prior conversation containing the questions. Let $P(w \mid C_i)$ be the probability of the word w occurring in the Q_i given context C_i. We define Informativeness as the conditional entropy:

$$H(Q_i \mid C_i) = -\sum_{w \in Q_i} P(w \mid C_i) \log P(w \mid C_i)$$

Conditional Entropy captures how much new information a follow-up question introduces relative to the prior questions in the conversation. A lower $H(Q_i \mid C_i)$ suggests redundancy amongst questions.

Relevance (Maxim of Relation): The Maxim of Relation emphasizes that follow-up questions should remain relevant to the ongoing conversation. A question that deviates significantly from the context can disrupt dialogue coherence.
We define the Relevance Score for the ith follow-up question Q_i, given its context C_i, as:

$$\text{Relevance Score}(Q_i, C_i) = \cos(v(Q_i), v(C_i))$$

where $v(Q_i)$ is the embedding of Q_i, and $v(C_i)$ is the average embedding of all previous questions:

$$v(C_i) = \frac{1}{|C_i|} \sum_{q_j \in C_i} v(q_j)$$

A higher cosine similarity indicates stronger contextual alignment, ensuring that follow-up questions contribute meaningfully to the conversation.

Clarity (Maxim of Manner): To evaluate Clarity, we use Average Dependency Distance (`ADD`), which measures how syntactically complex a sentence is. For each question Q_i, we define `ADD` as the average linear distance between words and their syntactic heads in the dependency tree. A lower `ADD` indicates a simpler, more comprehensible sentence structure. A well-formed follow-up question should be easy to understand and have a lower `ADD`. Shorter dependency distances indicate a syntactically simpler structure, making the question more direct and clear. In contrast, a higher `ADD` suggests a convoluted sentence, making comprehension harder.
We compute the Clarity score as follows:

$$\text{Clarity}(Q_i) = \frac{1}{1 + \texttt{ADD}(Q_i)}$$

[2] https://huggingface.co/FacebookAI/roberta-large-mnli

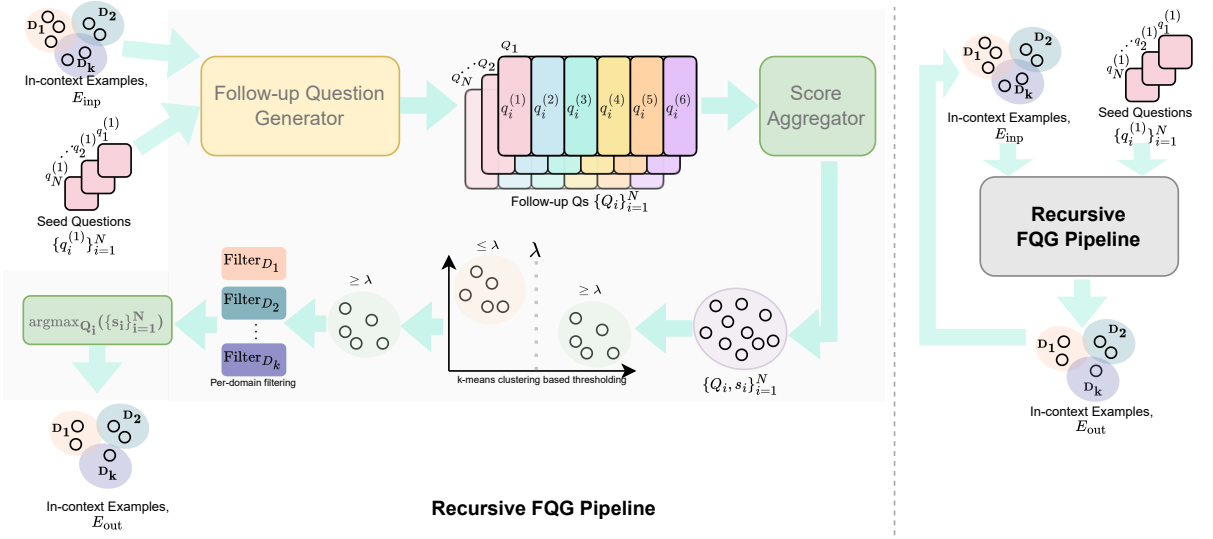

Figure 2: Recursive FQG pipeline. Starting from in-context examples $\mathcal{E}_{\text{inp}}$ drawn from domains $D_1, \ldots, D_k$ and 750 Level-1 (Remember) seed questions $\{q_i^{(1)}\}_{i=1}^{750}$, each seed is fed—via a few-shot prompt containing three human-annotated exemplars (seed + five follow-ups at Bloom Levels 2–6)—to an LLM-based Follow-up Question Generator. The model emits $M = 5$ candidates $\{q_i^{(j)}\}_{j=1}^{M}$, which are automatically scored on Logical Consistency, Informativeness, Relevance, and Clarity and aggregated into a single quality score. We apply K-means clustering with threshold λ to filter out low-quality sets and retain only those above λ. From this high-quality subset, we pick the top-scoring entry per domain to form a domain-diverse exemplar set, augment the prompt with these exemplars, and rerun the generator. Iterating this "generate → score → filter → cluster" loop yields the final out-of-domain examples $\mathcal{E}_{\text{out}}$.

2.2.2 LLM-based Reference-free Evaluation

Recent research highlights the potential of LLMs as reference-free evaluators for Natural Language Generation tasks (Chiang and Lee, 2023; Zheng et al., 2023; Liu et al., 2023). Building on this, we employed LLMs to evaluate follow-up questions based on four key metrics: *Logical Consistency, Informativeness, Relevance*, and *Clarity*, which are grounded in Gricean Maxims. The example of the evaluation prompts, structured following Siledar et al. (2024), are provided in Figure 6, 7, 8 & 9. For this evaluation, we used the `gpt-4o-mini` model.

2.3 B-FQG: Bloom's Taxonomy-based Follow-up Question Generation

We generate follow-up questions that progressively increase cognitive complexity according to Bloom's Revised Taxonomy (Appendix A.2), using 750 Level-1 (Remember) seed questions (see Figure 2). Each seed is input to an LLM-based Follow-up Question Generator via a few-shot prompt (Refer Figure 5 for the prompt) comprising three human-annotated examples, each consisting of a seed question and five follow-ups at Bloom Levels 2–6.

The LLM produces five follow-up questions per seed. We automatically score each complete set (Seed Question + 5 Follow-up Questions) on Logical Consistency, Relevance, Clarity, and Informativeness, aggregating these into a single quality score. We apply K-means clustering to these scores to define a threshold and retain only those entries above it.

From this high-quality subset, we select the top-scoring entry per domain to form a set of domain-diverse exemplars. We then augment the prompt with these exemplars and regenerate follow-ups for the lower-scoring seeds, repeating this bootstrap cycle until all entries meet our quality criteria. The follow-up questions were annotated to assess whether they adhered to the intended Bloom's levels, and it was found that they achieved an adherence accuracy of 96.68% (Table 5 in Appendix); for the full annotation guidelines, see Figure 10 (in Appendix).

3 Evaluation and Results

The quality of follow-up questions was evaluated using Grice's Cooperative Principle, a foundational theory in pragmatics that outlines how effective communication relies on adherence to four

conversational maxims: *Quality, Quantity, Relation,* and *Manner*. Each maxim offers valuable insights into the effectiveness and clarity of the follow-up questions in a conversational context.

This theoretical framework, based on Grice's maxims, provides a foundation for evaluating follow-up questions, guiding how they should function within a conversation to ensure logical consistency, appropriate informativeness, relevance, and clarity. We also conducted the human and LLM-based evaluations using the above metrics.

3.1 Human Evaluation

We evaluated a total of 375 follow-up questions generated from 75 randomly sampled seed questions. These questions were assessed by a human annotator on four metrics, which are rooted in Gricean Maxims. The result of the human evaluation is present in Table 2. We evaluated the follow-up questions generated by the Qwen2.5-7B-Instruct[3], Mistral-7B-Instruct[4], OLMoE-1B-7B-Instruct[5] and Llama-3.1-8B-Instruct[6] model across all four models scored above 4.4 out of 5 on every metric. Mistral-7B-Instruct achieved the highest logical consistency (4.79) and relevance (4.66), while Qwen-7B-Instruct led in informativeness (4.70) and clarity (4.79). The small differences in scores show that all four models generate consistently high-quality follow-up questions under the Gricean Maxims framework.

3.2 GriceWise Scores

Table 3 presents the evaluation of follow-up questions generated by different LLMs using GriceWise metrics (Section 2.2). Qwen-7B-Instruct (few-shot) achieved the highest scores in Logical Consistency, Relevance and Clarity. Mistral-7B-Instruct (few-shot) led in Informativeness. The performance gap between few-shot and zero-shot prompting reinforces the importance of in-context learning (Figure 4).

[3] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[4] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[5] https://huggingface.co/allenai/OLMoE-1B-7B-0924

[6] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

3.3 Validation of Automated Evaluation Methods

Table 6 (in Appendix) reports Spearman's ρ and Kendall's τ correlations between human judgments and two automated evaluation methods: GriceWise reference-free evaluation and LLM-based evaluation using `gpt-4o-mini`. GriceWise scores align moderately to strongly with human annotations ($\rho = 0.56$–0.72; $\tau = 0.47$–0.60), with Clarity showing the highest correspondence ($\rho = 0.72$; $\tau = 0.60$). LLM-based evaluation further improves these correlations ($\rho = 0.63$–0.76; $\tau = 0.62$–0.73), again peaking on Clarity ($\rho = 0.76$; $\tau = 0.73$). This confirms that both GriceWise and LLM-based methods reliably capture the same quality signals as human annotators.

3.4 Case Study

We evaluated both our seed and recursive follow-up questions' responses on a commercially deployed in-car AI assistant[7]. Table 4 shows the proportion of fallback responses, particularly the assistant's default "Didn't get that" reply, and how it varies across different cognitive levels. In a manual post-hoc annotation of the assistant's outputs, we found that,

1. Level 1 (Remember): 52% of questions were answered correctly, while the remaining 48% returned hallucinated content, generic/under-specified replies, or simple fallbacks ("Didn't get that," "Sorry, I don't have that information").

2. Level 6 (Create): Only 6% of questions were answered correctly; the other 94% produced hallucinations, generic responses, or fallback messages.

Such stark differences in response quality across cognitive levels highlight the pressing need to systematically recognize and address the limitations of the in-car AI assistant, especially given the high-stakes nature of in-vehicle interactions. With a correctness coverage as low as 6% at the highest cognitive level, there is a clear imperative to enhance the assistant's performance. This underscores the importance of integrating structured domain knowledge, such

[7] For confidentiality reasons, the specific car and in-car AI assistant names are not disclosed; we use "commercially deployed in-car AI assistant" instead.

Question Generation Model	Logical Consistency (↑)	Informativeness (↑)	Relevance (↑)	Clarity (↑)
Qwen-7B-Instruct	4.70	**4.70**	4.63	**4.79**
Mistral-7B-Instruct	**4.79**	4.65	**4.66**	4.78
OLMoE-1B-7B-Instruct	4.68	4.63	4.46	4.75
Llama-3.1-8B-Instruct	4.62	4.44	4.33	4.63

Table 2: Human evaluation scores (on a 5-point scale) for follow-up questions generated by four models—Qwen-7B-Instruct, Mistral-7B-Instruct, OLMoE-1B-7B-Instruct, and Llama-3.1-8B-Instruct—across four metrics: Logical Consistency, Informativeness, Relevance, and Clarity. Arrows next to each metric name indicate the scoring direction: (↑) denotes that higher scores are preferred.

Question Generation Models	Logical Consistency (↑)	Informativeness (↑)	Relevance (↑)	Clarity (↑)
Qwen-7B-Instruct	**0.9122**	0.5108	**0.6025**	**0.2743**
Mistral-7B-Instruct	<u>0.9052</u>	**0.5991**	<u>0.5917</u>	<u>0.2723</u>
OLMoE-1B-7B-Instruct	0.8720	0.4693	0.5569	0.2688
Llama-3.1-8B-Instruct	0.8893	<u>0.5906</u>	0.5559	0.2600

Table 3: Evaluation of Follow-up Question Generation Models on four metrics based on the GriceWise evaluation framework (Section 2.2). The best scores are bolded, and the second-best scores are underlined. Arrows next to each metric name indicate the scoring direction: (↑) denotes that higher scores are preferred.

as car manuals, and employing targeted prompt refinement strategies to improve the reliability and relevance of responses generated by LLM-powered in-car AI systems.

Level	% of Failure
1	45.33
2	12.00
3	45.33
4	10.67
5	17.33
6	26.67

Table 4: Proportion of fallback responses (e.g., "Didn't get that") from a commercially deployed in-car AI assistant across the six levels of Bloom's Taxonomy

4 Conclusion and Future Work

We presented a framework that leverages Bloom's Taxonomy to generate follow-up questions with increasing cognitive complexity. We employed Gricean-inspired evaluation metrics to assess the generated follow-up questions' logical consistency, informativeness, relevance, and clarity. Our human-annotated dataset, consisting of seed questions, was created adhering to Level 1 of Bloom's Taxonomy. Additionally, the follow-up questions were annotated by humans to confirm that 96.68% of the generated questions adhere to the cognitive levels. For future work, we plan to refine our evaluation metrics further and explore additional prompting strategies and model variations to enhance the follow-up question generation.

Limitations

Our approach is limited by the quality and scope of the human-annotated seed questions and the inherent capabilities of current LLMs. Due to confidentiality reasons, we could not mention the name of the in-car AI assistant we used to test our follow-up questions. Future work should extend human evaluation across a broader range of models and prompting strategies.

Ethics Statement

All human annotations were performed ethically with fair compensation. No personally identifiable information was used. Our data collection and annotation processes adhere to respecting privacy and fairness throughout the research.

Acknowledgements

The authors thank the anonymous reviewers for their constructive feedback, which helped improve this submission. We extend our sincere gratitude to the Computation for Indian Language Technology (CFILT) Lab at the Indian Institute of Technology Bombay for providing the computational resources that were indispensable for the successful completion of this research. We also extend our thanks to the annotators for their diligent and honest efforts.

References

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. *arXiv preprint arXiv:1705.00106*.

Mihail Eric and Christopher D Manning. 2017. Key-value retrieval networks for task-oriented dialogue. *arXiv preprint arXiv:1705.05414*.

Yubin Ge, Ziang Xiao, Jana Diesner, Heng Ji, Karrie Karahalios, and Hari Sundaram. 2023. What should I ask: A knowledge-driven approach for follow-up questions generation in conversational surveys. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 113–124, Hong Kong, China. Association for Computational Linguistics.

Bilal Ghanem, Lauren Lutz Coleman, Julia Rivard Dexter, Spencer McIntosh von der Ohe, and Alona Fyshe. 2022. Question generation for reading comprehension assessment by modeling how and what to ask. *arXiv preprint arXiv:2204.02908*.

Michael Heilman and Noah A Smith. 2010. Good question! statistical ranking for question generation. In *Human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics*, pages 609–617.

Thomas Huber and Christina Niklaus. 2025. Llms meet bloom's taxonomy: A cognitive view on large language model evaluations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5211–5246.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. Semantic graphs for generating deep questions. *arXiv preprint arXiv:2004.12704*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*.

Tejpalsingh Siledar, Swaroop Nath, Sankara Muddu, Rupasai Rangaraju, Swaprava Nath, Pushpak Bhattacharyya, Suman Banerjee, Amey Patil, Sudhanshu Singh, Muthusamy Chelliah, and Nikesh Garera. 2024. One prompt to rule them all: LLMs for opinion summary evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12119–12134, Bangkok, Thailand. Association for Computational Linguistics.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2024. Scieval: A multi-level large language model evaluation benchmark for scientific research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19053–19061.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Maria-Eleni Zoumpoulidi, Georgios Paraskevopoulos, and Alexandros Potamianos. 2024. Bloomwise: Enhancing problem-solving capabilities of large language models using bloom's-taxonomy-inspired prompts. *arXiv preprint arXiv:2410.04094*.

A Appendix

A.1 Grice's Maxims

Grice's Maxims are conversational principles proposed by Paul Grice to ensure effective communication. These maxims guide cooperative conversations and are categorized as follows:

- **Maxim of Quantity**: Provide as much information as necessary, but no more.
- **Maxim of Quality**: Be truthful; do not provide false information or unsupported claims.
- **Maxim of Relation**: Ensure relevance by staying on topic.
- **Maxim of Manner**: Be clear, brief, and orderly while avoiding ambiguity and obscurity.

These maxims help facilitate meaningful and effective communication by promoting clarity, relevance, and truthfulness in discourse.

Level 2	Level 3	Level 4	Level 5	Level 6
0.973	0.960	0.973	0.964	0.964

Table 5: Accuracy of generated questions across different levels of Bloom's taxonomy. Human annotator verified whether each question at a particular level followed the corresponding level of Bloom's taxonomy.

A.2 Bloom's Taxonomy

Bloom's Taxonomy (Figure 3) is a classification of learning objectives and skills that educators use to structure lessons, assessments, and learning outcomes. Originally proposed in 1956 by Benjamin Bloom, an educational psychologist at the University of Chicago, the taxonomy has been updated to include the following six levels of learning:

- **Remembering**: Retrieving, recognizing, and recalling relevant knowledge from long-term memory.
- **Understanding**: Constructing meaning from oral, written, and graphic messages through interpreting, exemplifying, classifying, summarizing, inferring, comparing, and explaining.
- **Applying**: Carrying out or using a procedure for execution or implementation.
- **Analyzing**: Breaking material into constituent parts and determining how the parts relate to one another and to an overall structure or purpose through differentiating, organizing, and attributing.
- **Evaluating**: Making judgments based on criteria and standards through checking and critiquing.
- **Creating**: Putting elements together to form a coherent or functional whole; reorganizing elements into a new pattern or structure through generating, planning, or producing.

This taxonomy provides a structured approach to designing curricula and assessments, ensuring a comprehensive learning experience.

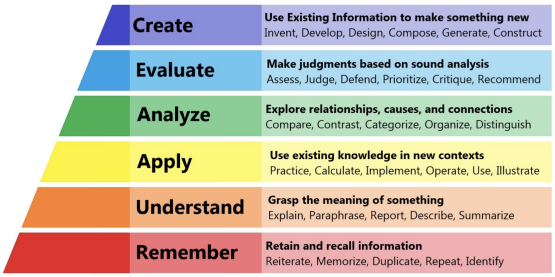

Figure 3: The Bloom's Taxonomy Pyramid: A hierarchical representation of cognitive learning levels, progressing from basic knowledge recall to complex creation and synthesis.

A.3 Example Follow-Up Questions for In-Car AI Assistants

Below is an example illustrating our multi-turn follow-up question generation for the call-making domain, demonstrating a progression in cognitive complexity based on Bloom's Taxonomy:

- **Seed Question (Level 1):** "How do I make a call?"
- **Follow-Up Question 1 (Level 2):** "What are the different options I have to make a call in this car?"
- **Follow-Up Question 2 (Level 3):** "How does the call-making process differ from my previous car model?"
- **Follow-Up Question 3 (Level 4):** "What are the advantages of using the car's built-in calling system over my phone's calling feature?"
- **Follow-Up Question 4 (Level 5):** "Can you explain how the car's calling system integrates with my phone's contact list and how it affects call quality?"
- **Follow-Up Question 5 (Level 6):** "How can I use the call-making feature in this car to improve my safety while driving, such as by using voice commands or hands-free modes?"

Evaluation Method	Logical Consistency (Maxim of Quality)		Informativeness (Maxim of Quantity)		Relevance (Maxim of Relation)		Clarity (Maxim of Manner)	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ
GriceWise Evaluation	0.57	0.48	0.56	0.47	0.61	0.52	0.72	0.60
LLM-based Evaluation	0.63	0.62	0.65	0.63	0.66	0.64	0.76	0.73

Table 6: Spearman's ρ and Kendall's τ correlation of human evaluation with GriceEise Evaluation and LLM-based evaluation across four metrics. `gpt-4o-mini` was used for LLM-based evaluation.

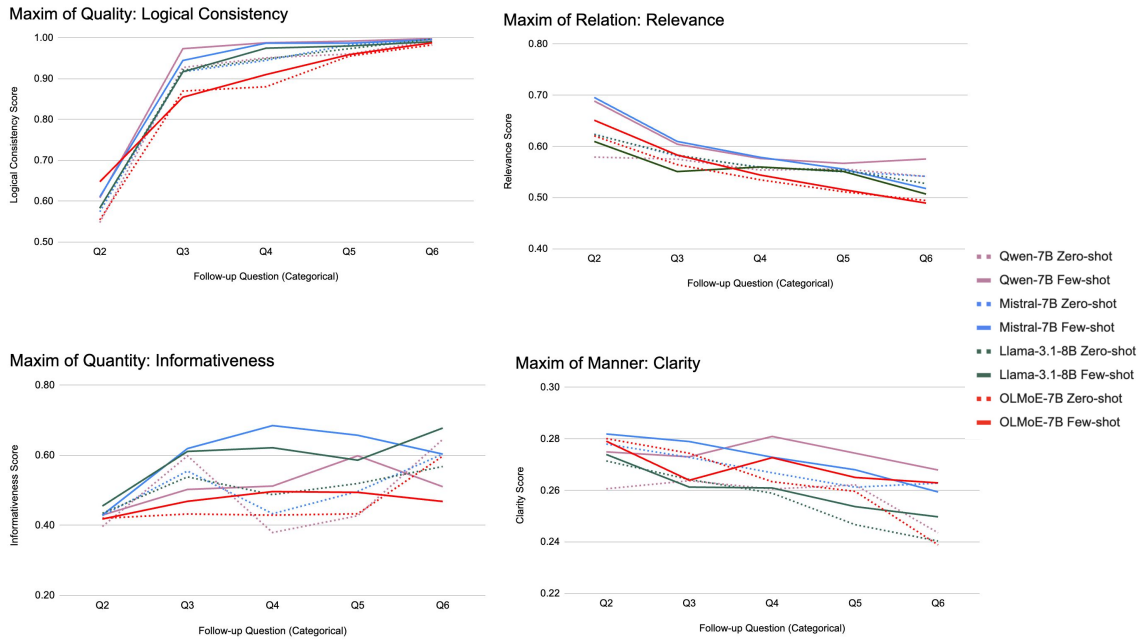

Figure 4: GriceWise (Logical Consistency, Informativeness, Relevance, Clarity) for different models and prompting strategies (zero-shot, few-shot) across follow-up questions Q2–Q6. Higher scores indicate stronger adherence to the respective maxim, capturing how well the model maintains coherence, relevance, informativeness, and clarity in follow-up question generation. Dotted lines represent zero-shot prompting and solid lines represent few-shot prompting.

Qualitative Insights on GriceWise Metric Trends:

- **Logical Consistency (Maxim of Quality)**: Sharp increase from Q2 to Q4, then plateaus; few-shot > zero-shot. The GriceWise evaluation for logical consistency is binary (0 or 1), so the sharp increase reflects a growing number of responses being judged fully consistent as the model gains context.
- **Informativeness (Maxim of Quantity)**: Gradual improvement across Q1 to Q5. Few-shot prompting provides better guidance, yielding richer follow-up questions.
- **Relevance (Maxim of Relation)**: Relevance gradually decreases from Q2 to Q6 as questions grow more abstract and harder to align with the main topic; few-shot prompting offers some improvement by providing better grounding, but cannot fully prevent the decline.
- **Clarity (Maxim of Manner)**: Clarity declines steadily as question chains grow longer, often introducing verbosity or ambiguity; few-shot examples help maintain concise and direct phrasing, mitigating this effect.

Task Description: You are an AI tasked with generating follow-up questions for a car driver to ask an in-car AI assistant. The questions will assess the AI's understanding of the car's features and design strictly based on the information provided in the seed question. The driver will begin with a Level 1 (Remember) question based on Bloom's Revised Taxonomy. Your task is to generate five follow-up questions corresponding to Levels 2 (Understand), 3 (Apply), 4 (Analyze), 5 (Evaluate), and 6 (Create), respectively. Each question should progress from simpler to more complex cognitive tasks.

Constraints:
Feature Neutrality: Do not assume, add, or imply any car features that are not explicitly mentioned or suggested in the seed question. Base all follow-up questions solely on the context given in the seed question.
Answer-Agnostic: Focus on the driver's interaction with the car and how the car's features enhance the driving experience without delving into internal technical details or making assumptions about additional features.
Driver-Focused Interaction: Ensure that all questions centre on the driver's use and experience with the car. Do not include questions regarding the car's internal mechanisms, data-acquisition methods, or any technical processes.
Single-Faceted: Each question must target a single concept or action to maintain clarity. Avoid compound or multi-part questions.
Sequential Progression: The follow-up questions should build upon each other, moving from basic recall (Level 1) to more advanced cognitive tasks (Level 6).
Bloom's Levels Only: Only generate questions for Levels 2 through 6 of Bloom's Revised Taxonomy. Do not introduce any levels beyond Level 6.

Explanation of Bloom's Revised Taxonomy Levels:
Level 1 (Remember): Involves recalling or recognizing facts and basic concepts. (This level is provided as the seed question.)
Level 2 (Understand): Involves explaining ideas or concepts. Questions at this level ask for clarification or interpretation.
Level 3 (Apply): Involves using information in new or concrete situations. Questions should prompt practical use or demonstration of how a feature could be used.
Level 4 (Analyze): Involves breaking information into parts and exploring relationships. Questions should prompt examination of reasons, causes, or underlying structures.
Level 5 (Evaluate): Involves making judgments based on criteria and standards. Questions should encourage assessment or justification of decisions.
Level 6 (Create): Involves putting elements together to form a new, coherent whole or proposing alternative solutions. Questions should prompt the generation of original ideas or new perspectives.

Input Format: <seed> seed_question_str </seed>
Output Format:
<question>question_1_str</question>
.
<question>question_5_str</question>

Instruction: Output only five lines, each corresponding to a question from level 2 to level 6 as described before, and nothing else. Do not provide any additional explanation or reasoning.

Figure 5: Prompt for Follow-up Question Generation based on Bloom's Taxonomy

Task Description: The purpose of evaluating questions based on the Maxim of Quality is to assess the truthfulness, accuracy, and reliability of the follow-up questions. Grice's Maxim of Quality suggests that communication should aim to be truthful and avoid saying anything that is false or for which the speaker lacks sufficient evidence. Evaluate whether the follow-up question maintains the integrity of the information provided by the previous question and whether it introduces any false, speculative, or unverifiable claims.

Evaluation Criteria: The task is to judge the extent to which the metric is followed by the follow-up question. Following are the scores and the evaluation criteria according to which scores must be assigned.
<score>1</score> - The metric is not followed at all while generating the follow-up question based on the previous questions.
<score>2</score> - The metric is followed only to a limited extent while generating the follow-up question based on the previous questions.
<score>3</score> - The metric is followed to a good extent while generating the follow-up question based on the previous questions.
<score>4</score> - The metric is followed mostly while generating the follow-up question based on the previous questions.
<score>5</score> - The metric is followed completely while generating the follow-up question based on the previous questions.

Metric: Maxim of Quality - For a follow-up question, it evaluate its alignment with the factual accuracy and truthfulness of the initial question. Consider whether the follow-up introduces any false, misleading, or speculative elements. Pay close attention to whether the question is rooted in facts and whether any claims made are verifiable. If the question is entirely accurate and grounded in truth, it should receive a higher score. If the question introduces errors, falsehoods, or speculative elements, it should receive a lower score.

Previous Questions:
{previous}

Follow-up Question:
{followup}

Evaluation Steps:
Follow the following steps strictly while giving the response:
1. First, write down the steps that are needed to evaluate the follow-up question as per the metric. Reiterate what metric you will be using to evaluate the follow-up question.
2. Give a step-by-step explanation if the follow-up question adheres to the metric, considering the previous questions as the input. Stick to the metric only for evaluation.
3. Next, evaluate the extent to which the metric is followed.
4. Rate the follow-up question using the evaluation criteria and assign a score within the <score></score> tags.

Note: Strictly give the score within <score></score> tags only e.g Score- <score>5</score>.
First, give a detailed explanation and then finally give a single score following the format: Score-<score>5</score>

Figure 6: Prompt for LLM-based evaluation of Maxim of Quality

Task Description: The purpose of evaluating questions based on the Maxim of Quantity is to assess whether the follow-up questions provide the appropriate amount of information. Grice's Maxim of Quantity suggests that communication should be as informative as is needed but not more than is required. The follow-up question should neither overwhelm with excessive detail nor leave important gaps in information. Assess whether the follow-up question is appropriately detailed or concise, neither under-informing nor over-informing.

Evaluation Criteria: The task is to judge the extent to which the metric is followed by the follow-up question. Following are the scores and the evaluation criteria according to which scores must be assigned.
<score>1</score> - The metric is not followed at all while generating the follow-up question based on the previous questions.
<score>2</score> - The metric is followed only to a limited extent while generating the follow-up question based on the previous questions.
<score>3</score> - The metric is followed to a good extent while generating the follow-up question based on the previous questions.
<score>4</score> - The metric is followed mostly while generating the follow-up question based on the previous questions.
<score>5</score> - The metric is followed completely while generating the follow-up question based on the previous questions.

Metric: Maxim of Quantity - For a follow-up question, it determines if the question is appropriately informative given the context of the conversation. Consider whether the question provides enough information to answer it or if it overcomplicates things by including irrelevant details. The perfect follow-up question will be balanced, providing enough context and detail to be clear and actionable without overwhelming the listener or leaving gaps. If the question provides the right amount of detail, score it higher. If it gives too little or too much, score it lower.

Previous Questions:
{previous}

Follow-up Question:
{followup}

Evaluation Steps:
Follow the following steps strictly while giving the response:
1. First, write down the steps that are needed to evaluate the follow-up question as per the metric. Reiterate what metric you will be using to evaluate the follow-up question.
2. Give a step-by-step explanation if the follow-up question adheres to the metric, considering the previous questions as the input. Stick to the metric only for evaluation.
3. Next, evaluate the extent to which the metric is followed.
4. Rate the follow-up question using the evaluation criteria and assign a score within the <score></score> tags.

Note: Strictly give the score within <score></score> tags only e.g. Score- <score>5</score>.

First, give a detailed explanation and then finally give a single score following the format: Score-<score>5</score>

Figure 7: Prompt for LLM-based evaluation of Maxim of Quantity

Task Description: The purpose of evaluating questions based on the Maxim of Relation is to assess the relevance of follow-up questions in relation to the preceding questions and the overall context. Grice's Maxim of Relation emphasizes that communication should be relevant and connected, meaning the follow-up question should logically follow from the previous question and maintain a coherent conversation. Assess whether the follow-up question is appropriately related to the previous question, both in terms of topic and context.

Evaluation Criteria: The task is to judge the extent to which the metric is followed by the follow-up question. Following are the scores and the evaluation criteria according to which scores must be assigned.
<score>1</score> - The metric is not followed at all while generating the follow-up question based on the previous questions.
<score>2</score> - The metric is followed only to a limited extent while generating the follow-up question based on the previous questions.
<score>3</score> - The metric is followed to a good extent while generating the follow-up question based on the previous questions.
<score>4</score> - The metric is followed mostly while generating the follow-up question based on the previous questions.
<score>5</score> - The metric is followed completely while generating the follow-up question based on the previous questions.

Metric: Maxim of Relation - It ensures that the follow-up question is relevant to the seed question and logically follows from the prior context. Look for continuity in the conversation's topic or subject matter; ensure the follow-up does not feel out of place or introduce unnecessary tangents. If the question feels disconnected or introduces unrelated ideas, it should receive a lower score. A highly relevant and contextually appropriate follow-up should receive a higher score.

Previous Questions:
{previous}

Follow-up Question:
{followup}

Evaluation Steps:
Follow the following steps strictly while giving the response:
1. First, write down the steps that are needed to evaluate the follow-up question as per the metric. Reiterate what metric you will be using to evaluate the follow-up question.
2. Give a step-by-step explanation if the follow-up question adheres to the metric, considering the previous questions as the input. Stick to the metric only for evaluation.
3. Next, evaluate the extent to which the metric is followed.
4. Rate the follow-up question using the evaluation criteria and assign a score within the <score></score> tags.

Note: Strictly give the score within <score></score> tags only e.g Score- <score>5</score>.
First, give a detailed explanation and then finally give a single score following the format: Score-<score>5</score>

Figure 8: Prompt for LLM-based evaluation of Maxim of Relation

Task Description: The purpose of evaluating questions based on the Maxim of Manner is to assess the clarity and conciseness of follow-up questions. Grice's Maxim of Manner suggests that communication should avoid ambiguity and be as clear and concise as possible to ensure that the listener can easily understand the message. Assess whether the follow-up questions adhere to these principles, focusing on how well the question conveys its intent and whether it does so in a straightforward and unambiguous manner.

Evaluation Criteria: The task is to judge the extent to which the metric is followed by the follow-up question. Following are the scores and the evaluation criteria according to which scores must be assigned.
<score>1</score> - The metric is not followed at all while generating the follow-up question based on the previous questions.
<score>2</score> - The metric is followed only to a limited extent while generating the follow-up question based on the previous questions.
<score>3</score> - The metric is followed to a good extent while generating the follow-up question based on the previous questions.
<score>4</score> - The metric is followed mostly while generating the follow-up question based on the previous questions.
<score>5</score> - The metric is followed completely while generating the follow-up question based on the previous questions.

Metric: Maxim of Manner - It considers whether the follow-up question can be understood easily in a first reading. Think about whether the question has any redundant parts that could be omitted. Ensure the wording is straightforward, and avoid complex sentence structures unless absolutely necessary. If the question feels awkward or the meaning seems unclear, lean towards giving it a lower score (1-3). If it's concise and the intent is immediately clear, it should score higher (4-5).

Previous Questions:
{previous}
Follow-up Question:
{followup}

Evaluation Steps:
Follow the following steps strictly while giving the response:
1. First, write down the steps that are needed to evaluate the follow-up question as per the metric. Reiterate what metric you will be using to evaluate the follow-up question.
2. Give a step-by-step explanation if the follow-up question adheres to the metric, considering the previous questions as the input. Stick to the metric only for evaluation.
3. Next, evaluate the extent to which the metric is followed.
4. Rate the follow-up question using the evaluation criteria and assign a score within the <score></score> tags.

Note: Strictly give the score within <score></score> tags only e.g Score- <score>5</score>.
First give a detailed explanation and then finally give a single score following the format: Score-<score>5</score>

Figure 9: Prompt for LLM-based evaluation of Maxim of Manner

Figure 10: Overview of the guideline which was used for data annotation for the seed questions.

These guidelines define how to frame and annotate follow-up questions for a car AI system. The goal is to ensure that the questions align with the car AI's capabilities and follow a structured approach based on Bloom's Taxonomy. This annotation task will work as a seed question for generating follow-up questions.

General Principles

1. **Action or Information Focus**: For POI (Point of Interest)or navigation tasks, focus on recalling details like location, route, or destination.
2. **Task-Oriented and Contextual**: Ensure that the questions are actionable, focusing on what the car AI can recall about POIs, weather, or time-related queries.
3. **Simple, Direct Questions**: Ask specific, factual questions that a driver would need to recall or verify to continue their task, such as routes, locations, or specific information like weather or time.
4. **Avoid Redundancy**: Do not ask for general or already known information (e.g., "Who do I want to call?"). Instead, focus on recalling detailed, task-specific information that will aid in decision-making.
5. **Driver-Centric Questioning: Annotators should frame questions as if they are a car driver interacting with an in-car AI chatbot.**

Domain-Specific Guidelines

Phone Domain

Imperative to Interrogative Transformation: Avoid forced interrogative conversions. Instead, structure questions naturally.

Bloom's Level 1 (Remembering/Recall)

What, Which, How

Commands & Interrogative Conversions:

Command	How	What	Which
Call	How can I make a call?	What is the command to make a call?	How do I make a call?
Call	How do I call John Smith?	What is the command to call John Smith?	Which number will be dialled if I say 'Call John Smith'?
Dial <012-345-7890>	How do I dial the number 012-345-7890?	What is the command to dial the number 012-345-7890?	How do I dial a number manually?
Change Bluetooth Device	How do I change the Bluetooth device?	What is the command to change the Bluetooth device?	Which device is currently connected via Bluetooth?

Send Message

Commands & Interrogative Conversions:

Command	How	What
Send Message	How do I send a message?	What is the command to send a message?
Send Message to	How do I send a message to John Smith?	What is the command to send a message to John Smith?

Weather Queries

How - Condition-based recall

- *How is the weather today?*
- *How was the weather yesterday in Hyderabad?*
- *How is the weather next Sunday in Hyderabad?*

What - Detail-based recall

- *What is the temperature today?*
- *What was the highest temperature yesterday?*

Which - Comparison-based recall

- *Which city had the highest temperature yesterday?*

Date and Time Queries

What - Factual recall

- *What time is it in Tokyo?*
- *What is the date today?*

How - Quantity-based recall

- *How many days are there between today and March 3rd?*

When - Time-based recall

- *When is Diwali?*

Which - Comparison-based recall

- *Which time zone does Tokyo follow?*

Radio Control

What - Information recall

- *What is the current radio station?*

How - Task recall

- *How do I tune to FM 100.1?*

Which - Option selection

- *Which AM station can I switch to?*

Media Control

What - Status recall

- *What media is currently playing?*

How - Task recall

- *How do I turn off the media?*
- *How do I turn off Bluetooth audio?*

Is - Status check

- *Is the media turned off?*
- *Is the Bluetooth turned on?*

NLU Commands

What - Information recall

- *What is the current temperature?*
- *What is the condition of the windows?*

How - Task recall

- *How do I clear the fog on the windshield?*
- *How do I adjust the windows?*

Can - Feasibility check

- *Can I cool down the car?*
- *Can I clear the fog on the windshield?*

Automatic Temperature Control

What - Status recall

- *What is the current fan speed?*

How - Task recall

- *How do I activate the front defroster?*

Can - Feasibility check

- *Can I open the sunroof?*

Is - Status check

- *Is the climate control on?*
- *Is the air conditioning on?*