

DRIVE: Disfluency-Rich Synthetic Dialog Data Generation Framework for Intelligent Vehicle Environments

Anshul Chavda¹ M Jagadeesh² Chintalapalli Raja Kullayappa² B Jayaprakash²
Medchalimi Sruthi² Pushpak Bhattacharyya¹

¹Indian Institute of Technology Bombay ²Hyundai Motors India Engineering
{anshulchavda, pb}@cse.iitb.ac.in

Abstract

In-car conversational AI is becoming increasingly critical as autonomous vehicles and smart assistants gain widespread adoption. Yet, existing datasets fail to capture the spontaneous disfluencies—hesitations, false starts, repetitions, and self-corrections—that characterize real driver–AI dialogs. To address this, we introduce DiscoDrive, a synthetic corpus of 3,500 multi-turn dialogs across seven automotive domains, generated using a two-stage, prompt-driven pipeline that dynamically integrates disfluencies during synthesis. We show that DiscoDrive is effective both as a training resource—enabling DialoGPT-Medium and T5-Base to match or exceed KVRET-trained models on the MultiWOZ 2.2 and Schema-Guided Dialogue (SGD) relevant test sets (BLEU-4 +0.26–0.61; METEOR +2.10; ROUGE-L +3.48; BERTScore F_1 +1.35–3.48)—and as a data augmentation resource in low-resource scenarios, delivering additional gains of up to BLEU-4 +0.38, METEOR +1.95, ROUGE-L +2.87, and BERTScore F_1 +4.00 when combined with 10% of KVRET. Human evaluations further confirm that dialogs sampled from DiscoDrive are rated higher than KVRET’s human-collected dialogs in naturalness (3.8 vs. 3.6) and coherence (4.1 vs. 4.0), and are perceived as more context-appropriate than leading post-hoc methods (e.g., LARD), without compromising clarity. DiscoDrive thus fills a critical gap in existing resources and serves as a versatile corpus for both training and augmenting conversational AI, enabling robust handling of real-world, disfluent in-car interactions.

1 Introduction

Synthetic dialog data generation has become a key area in conversational AI, enabling the creation of large-scale datasets for model training when real-world data is scarce or expensive to collect. Despite significant progress in generating high-quality dialogs for general-purpose assistants like Siri or

Alexa, specialized datasets for domains such as automotive applications remain limited. Existing resources, such as KVRET (Eric et al., 2017), focus on fluent, task-oriented dialogs across limited domains (e.g., navigation and weather) and do not capture the diversity or spontaneous characteristics needed for more complex, domain-specific scenarios—particularly those involving safety, diagnostics, or urgent requests. This gap is critical, as in-car AI systems must manage unique communication dynamics under high-pressure conditions.

Another key issue is the inherent fluidity of spoken language, which often includes various errors and self-corrections made on the fly. When speakers become aware of an error, they naturally tend to modify, rephrase, or even completely restart their speech. These spontaneous interruptions and self-corrections—commonly referred to as disfluencies (Shriberg, 1994)—are a hallmark of natural conversation. Numerous theoretical studies have investigated the patterns of these disfluencies and how they affect human communication (Sparks, 1995; Shriberg, 1994; Plejert, 2004; Colman and Healey, 2011; Emrani and Hooshmand, 2019).

In natural, real-time spoken interactions—such as those in driver-AI systems—disfluencies are inevitable. Studies in spontaneous speech (Zayats et al., 2019) and research in naturalistic settings (Seyfeddinipur, 2006) confirm that fillers (e.g., “uh” and “um”), false starts, and self-repairs frequently occur even under high cognitive load. Despite their prevalence, disfluencies remain underrepresented in task-specific dialog datasets. For instance, the widely used Switchboard corpus contains only 5.9% of tokens annotated as disfluent (Charniak and Johnson, 2001), limiting exposure to the full range of disfluency phenomena encountered in driver-AI interactions. Moreover, disfluencies disrupt the regular, fluent structure that language models are typically trained on. Research by Ferreira and Bailey (2004) shows that such interruptions can

impair syntactic parsing and semantic integration, while [Fox Tree \(2001\)](#) demonstrate that filler words carry subtle prosodic cues reflecting planning delays—factors that further challenge the robustness of conversational AI.

Consequently, models fine-tuned on fluent, task-oriented datasets may struggle to interpret and respond effectively to the fragmented, disfluent inputs intrinsic to automotive contexts. Addressing this shortfall requires the development of domain-specific synthetic dialog datasets that not only replicate the natural flow of driver-AI interactions but also incorporate a rich variety of disfluencies, ultimately enhancing the naturalness and adaptability of AI systems in real-world automotive environments.

To address the limitations of existing dialog datasets and the unique requirements of human-car AI interactions, we present DiscoDrive (**DI**sfluency-Enriched **S**ynthetic **CO**nversations for **DRIVE**r-Car AI), a synthetic dialog dataset, with the following contributions:

1. **DiscoDrive**, a comprehensive dataset of 3,500 dialogs spanning seven domains, each with varying dialog lengths to reflect real-world in-car AI interactions. On average, each dialog contains 8 turns, balancing brief and complex conversations. (Section 3)
2. **Dynamic disfluency integration** directly during dialog generation via advanced prompt-driven generative modeling, **eliminating the need for post-hoc disfluency addition**. (Section 3)
3. **Downstream evaluations** on relevant test subsets of multiWOZ 2.2 ([Zang et al., 2020](#)) & SGD ([Rastogi et al., 2020](#)), finding that DialoGPT-Medium and T5-Base fine-tuned solely on DiscoDrive corpus achieve absolute BLEU-4 gains of **(+0.26 – +0.61)** and BERTScore F_1 improvements of **(+1.35 – +3.48)** over models fine-tuned on KVRET’s training set, while also surpassing them on ROUGE-L and METEOR, **validating its use as an alternative resource for training in-car conversational AI**. (Section 5).
4. In **low-resource scenarios**, augmenting scarce human data with our DiscoDrive yields substantial uplifts—e.g., BLEU-4 increases by +0.38 and BERTScore F_1 by (+3 – +4)

over real-only baselines, **confirming that DiscoDrive is an effective data augmentation resource in low-resource settings** (Section 5).

5. **Extensive human evaluations** reveal that DiscoDrive achieves significantly **higher ratings in disfluency naturalness and appropriateness compared to a leading post-hoc insertion technique** ([Passali et al., 2022](#)), without compromising clarity. Furthermore, it demonstrates **improvements in naturalness, human-likeness, and lexical diversity compared to the KVRET** ([Eric et al., 2017](#)) baseline (Section 5).

2 Related Work

2.1 Synthetic Dialog Data Generation

Traditional methods have relied on rule-based systems and sequence-to-sequence models, often utilizing manually curated datasets that lack the spontaneous and diverse characteristics of natural conversations. In contrast, recent advances in pre-trained language models like GPT-3 ([Brown et al., 2020](#)), GPT-4o ([OpenAI, 2024](#)), and DialoGPT ([Zhang et al., 2020](#)) have enabled the generation of more coherent, contextually rich conversations.

Newer frameworks such as PLACES ([Chen et al., 2023](#)) and semi-automated approaches like those presented by [Shah et al. \(2018\)](#) have further advanced synthetic dialog generation. In addition, recent works have specifically targeted complex conversational phenomena: TOAD ([Liu et al., 2024](#)) and LUCID ([Stacey et al., 2024](#)) generate task-oriented dialogs with diverse response styles and intricate utterance structures, while PRESTO ([Goel et al., 2023](#)) introduces explicitly labeled, multilingual dialogs that capture realistic conversational dynamics.

2.2 Disfluency Handling in NLP

Disfluencies, such as repetitions, false starts, and filler words, are common in spontaneous speech but challenging to model in NLP. Early approaches focused on detection and removal using statistical methods like hidden Markov models and rule-based systems ([Hough and Schlangen, 2015](#)). With advances in deep learning, models such as bidirectional LSTMs and transformers have achieved better accuracy in detecting disfluencies ([Zayats et al., 2016](#)).

While progress has been made in detection, less emphasis has been placed on generating disfluencies for training dialog systems. Recent work, such as (Yang et al., 2020), has explored disfluency generation for data augmentation, while (Passali et al., 2022) demonstrated large-scale disfluency generation to enhance model robustness. Additionally, (Marie, 2023) presented methods for generating disfluent dialog to improve the resilience of conversational systems to unstructured user input.

3 Methodology

Our approach to generating synthetic dialogs for driver-AI interactions leverages a two-step process involving conversation scenario generation and dialog simulation. We utilize advanced language models to simulate natural conversational dynamics, including disfluencies, while covering a range of realistic automotive scenarios. The methodology includes turn-based prompting, dynamic disfluency integration, and multi-domain coverage, all aimed at producing a high-quality dataset that reflects the complexities of real-world driver-AI interactions.

3.1 Two-Step Generation Pipeline

The synthetic dialog data generation pipeline is structured into two main stages:

Step 1: Conversation Scenario Generation.

We generated 500 diverse conversation scenarios for each domain using GPT-4o (OpenAI, 2024). These scenarios simulate realistic situations for dialogs within the automotive domain, such as "The driver wants to find the shortest route from Mumbai to Pune, avoiding traffic and toll roads." in the Navigation domain. To guide GPT-4o (OpenAI, 2024), we employed few-shot learning with a curated set of 10–20 human-written examples per domain, ensuring contextual relevance and diversity. The resulting scenarios provide the foundational context for the subsequent dialog simulation stage.

Step 2: Dialog Simulation. Using the Llama-3.1-8B-instruct model (MetaAI, 2024), we simulated multi-turn dialogs between the driver and the car AI. The simulation alternates between the two roles, with prompts tailored to each speaker’s conversational style—informal and disfluent for the driver, concise and task-focused for the car AI. Conversation history was limited to the last six exchanges to maintain coherence while balancing computational efficiency. This process ensured that the generated dialogs were contextually appropri-

ate and aligned with the scenario.

3.2 Turn-Based Prompting for Dialog Simulation

The dialog simulation process utilized a turn-based framework, alternating between prompts for the driver and the car AI to create realistic and contextually coherent exchanges.

Driver Prompts: Driver prompts were crafted to emulate the characteristics of natural communication in driving scenarios. These prompts simulated scenarios where the driver might be multitasking or distracted, incorporating informal language and natural disfluencies such as hesitations, repetitions, and self-corrections. This approach ensured that the generated dialogs reflected the spontaneity and variability of real-world interactions.

Car AI Prompts: The car AI prompts were designed to produce responses that are concise, informative, and aligned with the driver’s requests. The tone was deliberately conversational and friendly, reflecting the role of an in-car assistant, while prioritizing clarity and efficiency in addressing the driver’s needs.

At each turn, the model generated text based on the evolving conversation history, allowing it to leverage prior exchanges to maintain coherence and contextual relevance. Dynamic adjustments were made to the prompts during the dialog generation process, tailoring them to the specific stage of the conversation. For regular turns, prompts ensured a natural flow of exchanges, while for concluding turns, prompts guided the model to wrap up the dialog concisely and contextually. All prompt templates used to generate DiscoDrive are provided in Appendix A.3.

3.3 Dynamic Disfluency Integration

We dynamically integrated disfluencies during the generation process via prompt-driven generative modeling. This approach is motivated by extensive linguistic research demonstrating that natural speech is inherently disfluent.

For instance, repetitions frequently occur when speakers hesitate or emphasize a point, reflecting underlying cognitive load or uncertainty (Shriberg, 1994). Similarly, false starts and subsequent corrections are common in spontaneous conversation, as speakers often self-edit mid-utterance to clarify their intended message (Sparks, 1995; Plejert, 2004; Colman and Healey, 2011). Filler words like “um” and “uh” are pervasive in everyday speech

No.	Type	Description	Example
1	Repetitions	Repeating a word or phrase to simulate hesitation or emphasize a point.	"I think, I think we should take the next exit."
2	False Starts	Beginning a thought but switching direction mid-sentence, reflecting spontaneous speech corrections.	"We could—actually, let’s try the other route."
3	Filler Words	Incorporating words like "um" or "uh" to represent pauses in thinking or uncertainty.	"Can you, um, check the tire pressure?"
4	Pauses	Using ellipses ("...") to indicate brief pauses or hesitation in speech.	"I think we’ll be there... um, soon."
5	Corrections	Revising a previously stated idea or instruction to reflect common human speech adjustments.	"Turn left—no, wait, I mean right."

Table 1: Categories of Disfluencies Used in Dialog Generation: Definitions and Examples.

and serve as markers of planning delays or uncertainty, providing listeners with subtle prosodic cues (Fox Tree, 2001; Ferreira and Bailey, 2004). Additionally, natural pauses—represented in transcripts by ellipses—occur as speakers momentarily pause to organize their thoughts.

The types and examples of these disfluencies are detailed in Table 1. By incorporating these empirically validated disfluency types, our synthetic dialogs more accurately capture the spontaneous, variable nature of real-world driver–AI interactions, thereby enhancing the robustness and adaptability of conversational AI systems.

3.4 Domains and Scenarios

To ensure comprehensive coverage of automotive-related interactions, our dataset spans seven distinct domains:

1. **Navigation:** Involving requests for directions, traffic updates, and finding routes.
2. **Car Maintenance and Diagnostics:** Addressing vehicle status checks, troubleshooting, and scheduling services.
3. **Safety and Emergency Assistance:** Focusing on alerts, emergency handling, and safety-related inquiries.
4. **Entertainment:** Dialogs about media controls, music, radio, and other entertainment options.
5. **Local and On-Route Attractions and Activities:** Providing information on nearby points of interest, restaurants, and events.

6. **Car Functions:** Interactions about vehicle controls, such as adjusting air conditioning or seat settings.

7. **Weather:** Queries related to current weather conditions, forecasts, and weather-related alerts.

For each domain, we generated 500 unique conversation scenarios using GPT-4o (OpenAI, 2024), resulting in a total of 3,500 dialogs. The dialogs vary in length (6, 8, 10, 12, 14 turns) to reflect the range of real-world driver-AI interactions, from brief task-oriented exchanges to more complex, context-rich discussions. This range balances realism and diversity, ensuring the dataset captures the practical nuances of automotive conversations while supporting robust model training.

4 Experimental Setup

4.1 Compute Infrastructure

We ran generation on NVIDIA A40 and A100 GPUs (48 GB and 80 GB VRAM) using bfloat16 precision. Scenario and dialog synthesis consumed 40 GPU-hours; downstream fine-tuning (DialogPT-Medium (Zhang et al., 2020), T5-Base (Raffel et al., 2020)) and inference together used 30 GPU-hours.

4.2 Software Stack

Our pipeline is implemented in Python with HuggingFace Transformers v4.x. We use GPT-4o (OpenAI, 2024) for scenario generation and LLaMA-3.1-8B-Instruct (MetaAI, 2024) for simulation. Downstream models (DialogPT-Medium (Zhang et al., 2020), T5-Base (Raffel et al., 2020)) are

fine-tuned via HuggingFace’s Seq2SeqTrainer and Trainer APIs, with dynamic padding, gradient accumulation (batch size 16), and AdamW (lr = 5e-5 for DialoGPT-Medium (Zhang et al., 2020), 3e-5 for T5-Base (Raffel et al., 2020)) to ensure efficient multi-GPU training.

5 Results

This section presents the results and key observations from our experiments and evaluations.

5.1 Automatic Evaluation

To evaluate the lexical richness of our DiscoDrive dataset compared to KVRET dataset (Eric et al., 2017), we employed N-distinct (Li et al., 2016), a widely-used metric for assessing lexical diversity. The results of the N-Distinct evaluation are presented in Table 2.

N-Gram	KVRET	DiscoDrive Dataset
1-gram	0.0109	0.0124
2-gram	0.1040	0.1234
3-gram	0.2902	0.3428
4-gram	0.4808	0.5425

Table 2: N-Distinct scores for lexical diversity in KVRET (Eric et al., 2017) and DiscoDrive.

Observations: DiscoDrive achieves consistently higher N-Distinct scores across all n-gram levels compared to KVRET (Eric et al., 2017), demonstrating higher lexical diversity.

However, it is important to note that lexical diversity alone does not capture other essential aspects of dialog quality, such as naturalness and coherence, which are addressed through human evaluations in this work.

5.2 Human Evaluation

To evaluate the quality of DiscoDrive, two human evaluators performed individual dialog assessments, pairwise comparisons, and a targeted evaluation of dynamic versus post-hoc disfluency integration. For these evaluations, we used stratified subsets of 140 dialogues each from the DiscoDrive and KVRET datasets (Eric et al., 2017), ensuring equal representation across domains. A detailed description of the human evaluation setup can be found in Section A.1.

5.2.1 Intrinsic Evaluation Results

Human evaluators rated the dialogs based on various criteria using a 5-point Likert scale. The results

are summarized in Table 3.

Metric	KVRET	DiscoDrive
Naturalness	3.6 (± 0.18)	3.8 (± 0.18)
Coherence	4.0 (± 0.17)	4.1 (± 0.16)
Engagement	4.0 (± 0.17)	3.8 (± 0.17)
Consistency	4.3 (± 0.15)	4.2 (± 0.16)
On-topic	4.9 (± 0.04)	4.7 (± 0.10)

Table 3: Intrinsic evaluation results based on a 5-point Likert scale. The values represent the average scores from two human evaluators and the confidence intervals.

Observations: DiscoDrive dataset outperforms KVRET (Eric et al., 2017) in naturalness (3.8 vs. 3.6) and coherence (4.1 vs. 4.0), highlighting the positive impact of dynamically integrated disfluencies on dialog realism. However, the slight drop in engagement (3.8 vs. 4.0) and on-topic relevance (4.7 vs. 4.9) reflects the inherent variability introduced by disfluencies, which can occasionally detract from task focus. Despite these trade-offs, the synthetic dataset maintains strong consistency (4.2 vs. 4.3), demonstrating its suitability for training models that balance conversational realism and task-oriented performance.

5.2.2 Comparative Evaluation Results

In pairwise comparisons, evaluators selected which dialog was better across various criteria. The results are visualized in Figure 1.

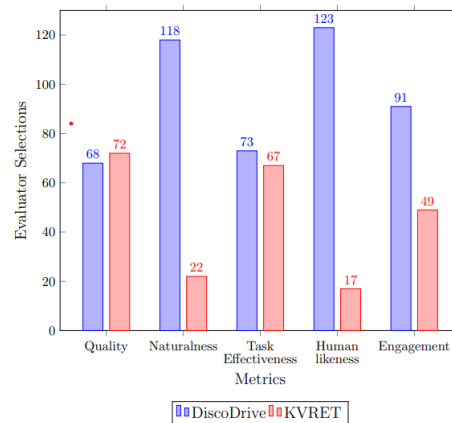


Figure 1: Comparative evaluation results for KVRET and DiscoDrive. The values represent the average scores from two human evaluators.

Observations: The DiscoDrive dataset outperformed KVRET (Eric et al., 2017) in naturalness (118 vs. 22), human-likeness (123 vs. 17), and engagement (91 vs. 49), demonstrating the success

of dynamically integrated disfluencies in producing more human-like, spontaneous dialogs.

While KVRET showed a slight edge in overall quality (72 vs. 68), likely due to its structured design, the synthetic dataset remained competitive in task effectiveness (73 vs. 67), balancing realism and task relevance effectively.

These results highlight the trade-off between conversational variability and structure but affirm the DiscoDrive’s strength in delivering realistic and engaging dialogs.

5.2.3 Evaluation of Dynamic vs. Post-Hoc Disfluency Integration

To further evaluate the impact of the disfluency integration approach, we applied the LARD method (Passali et al., 2022) for post-hoc insertion of disfluencies to the driver utterances in the KVRET (Eric et al., 2017) dialogs. Using the same subset of 140 dialogs employed in our other human evaluations, two human evaluators rated the dialogs on a 5-point Likert scale for various metrics. The results are summarized in Table 4.

Metric	KVRET	DiscoDrive
Naturalness	3.6 (± 0.17)	4.2 (± 0.16)
Appropriateness	3.4 (± 0.18)	4.3 (± 0.15)
Clarity	4.2 (± 0.16)	4.0 (± 0.16)

Table 4: Evaluation results for KVRET dialogs with post-hoc disfluency insertion using LARD (Passali et al., 2022) and DiscoDrive dialogs with dynamic disfluency integration. The values represent the average scores from two human evaluators and the confidence intervals.

Observations: The results indicate that dialogs generated with dynamic disfluency integration are perceived as significantly more natural and contextually appropriate than those modified using the LARD method (Passali et al., 2022). While clarity remains comparable between the two methods, the marked improvements in naturalness (4.2 vs. 3.6) and appropriateness (4.3 vs. 3.4) validate our approach for enhancing the realism of driver-AI interactions. Examples of post-hoc disfluency insertions using LARD method (Passali et al., 2022) are provided in Appendix A.2.

5.3 Experimental Evaluation of Full-Data Fine-Tuning for Car AI Response Generation

We assess the utility of our DiscoDrive dataset as a complete training resource by fine-tuning two standard dialogue models—DialoGPT-Medium (Zhang

et al., 2020) and T5-Base (Raffel et al., 2020)—under three conditions: (i) zero-shot (no fine-tuning), (ii) fine-tuned on the full KVRET training split (2,424 dialogs), and (iii) fine-tuned on our full DiscoDrive corpus (3,500 dialogs). We evaluate on dialog subsets (220 dialogs each) curated from multiWOZ 2.2 (Zang et al., 2020) and SGD (Rastogi et al., 2020) by filtering for services aligned with in-car assistant functionality, including domains such as navigation, weather updates, hotel reservations, attraction search, and restaurant inquiries. Evaluation metrics include BLEU-1–4, ROUGE-L, METEOR, and BERTScore F_1 .

multiWOZ 2.2 Results: Table 5 shows that models trained on the DiscoDrive corpus consistently outperform their KVRET-trained counterparts across BLEU, ROUGE, METEOR, and BERTScore. For DialoGPT (Zhang et al., 2020), fine-tuning on DiscoDrive yields a 21× improvement in BLEU-4 compared to KVRET (Eric et al., 2017) (0.64 vs. 0.03), and a +2.5 point increase in BERTScore F_1 (85.20 vs. 82.71). T5-Base (Raffel et al., 2020) shows a similar trend: although KVRET (Eric et al., 2017) leads slightly on BLEU-1, the DiscoDrive-trained model performs better on higher-order BLEU (BLEU-2–4), METEOR ($\uparrow +13.5$ points), ROUGE-L ($\uparrow +3.1$), and BERTScore F_1 ($\uparrow +1.35$).

SGD Results: Table 6 reports that the performance gap is even more pronounced. DialoGPT (Zhang et al., 2020) fine-tuned on DiscoDrive achieves a BLEU-4 of 0.28, substantially outperforming the KVRET-trained model (0.02), alongside a +3.5 point gain in BERTScore F_1 . For T5-Base (Raffel et al., 2020), DiscoDrive fine-tuning yields the strongest results across all metrics except BLEU-1, including +0.59 BLEU-4 and +1.72 BERTScore F_1 over the KVRET-trained version.

Analysis: Across both multiWOZ 2.2 (Zang et al., 2020) and SGD subsets (Rastogi et al., 2020), and for both DialoGPT-Medium (Zhang et al., 2020) and T5-Base (Raffel et al., 2020), fine-tuning on the DiscoDrive corpus produces generation quality that consistently surpasses that achieved using the human-curated KVRET dataset (Eric et al., 2017) (Tables 5–6). These results confirm the robustness and versatility of our synthetic dialogs, demonstrating DiscoDrive as a viable, scalable alternative to real-world annotations for comprehensive, full-scale training and rapid domain adapta-

Model & Configuration	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	BERTScore F ₁
<i>DialoGPT-Medium</i>							
Zero-Shot	01.10	00.14	00.05	00.01	3.64	2.24	74.82
KVRET Fine-Tuned	05.55	0.51	0.14	0.03	9.41	6.46	82.71
Synthetic Fine-Tuned	14.10	3.18	1.39	0.64	16.11	20.51	85.20
<i>T5-Base</i>							
Zero-Shot	12.34	2.10	0.95	0.42	10.20	12.80	78.30
KVRET Fine-Tuned	18.30	2.90	1.23	0.61	13.05	9.28	84.13
Synthetic Fine-Tuned	16.28	3.85	1.64	0.77	16.19	22.78	85.48

Table 5: Full-data fine-tuning results on the multiWOZ 2.2 in-car subset (beam=5). DiscoDrive-only fine-tuning outperforms KVRET on nearly all metrics.

Model & Configuration	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	BERTScore F ₁
<i>DialoGPT-Medium</i>							
Zero-Shot	00.82	00.09	00.02	00.00	2.90	1.94	75.08
Fine-Tuned on KVRET	01.91	0.18	0.06	0.02	7.03	5.21	81.23
Fine-Tuned on DiscoDrive	9.98	1.18	0.68	0.28	12.35	17.57	84.71
<i>T5-Base</i>							
Zero-Shot	08.28	01.39	00.56	00.25	8.60	14.16	82.98
Fine-Tuned on KVRET	15.39	2.86	1.31	0.67	10.51	8.03	83.83
Fine-Tuned on DiscoDrive	14.47	3.91	1.86	0.90	15.28	22.54	85.55

Table 6: Full-data fine-tuning results on the SGD in-car subset (beam=5). DiscoDrive-only fine-tuning outperforms KVRET on nearly all metrics.

tion in conversational AI.

5.4 Experimental Evaluation of Low-Resource Data Augmentation for Car AI Response Generation

To evaluate the utility of our DiscoDrive dataset in low-resource scenarios, we fine-tune DialoGPT-Medium (Zhang et al., 2020) and T5-Base (Raffel et al., 2020) under two configurations: (i) using only 10% of KVRET’s training split (242 dialogs), and (ii) augmenting the same 10% KVRET subset with the full DiscoDrive corpus (3,500 dialogs). We evaluate on the same dialog subsets (220 dialogs each) curated from multiWOZ 2.2 (Zhang et al., 2020) and SGD (Rastogi et al., 2020) as used in the full-data experiments. Evaluation metrics include BLEU-1–4, ROUGE-L, METEOR, and BERTScore F₁. Table 7 & 8 presents the low-resource results.

multiWOZ 2.2 Results: As reported in Table 7, DiscoDrive augmentation substantially improves generation quality across both models. For DialoGPT-Medium (Zhang et al., 2020), adding synthetic dialogs increases BLEU-4 from 0.05 to 0.58 and boosts BERTScore F₁ from 81.96 to 85.40—narrowing the gap toward full-data performance. For T5-Base (Raffel et al., 2020), we observe similar gains in higher-order BLEU, METEOR (↑+14.6), ROUGE-L (↑+7.6), and BERTScore F₁ (↑+2.4). These gains validate

the complementarity of synthetic disfluency-rich dialogs in enhancing low-data generalization.

SGD Results: Table 8 reveals analogous gains on the SGD in-car subset: DialoGPT-Medium (Zhang et al., 2020) achieves a BLEU-4 increase from 0.00 to 0.88 and a +4.9-point rise in BERTScore F₁, while T5-Base (Raffel et al., 2020) sees its BLEU-4 jump from 1.06 to 1.26 and records METEOR and ROUGE-L improvements exceeding +10 points each.

Analysis: Across both evaluation sets, multiWOZ 2.2 (Zhang et al., 2020) and SGD (Rastogi et al., 2020)—and for both DialoGPT-Medium (Zhang et al., 2020) and T5-Base (Raffel et al., 2020), augmenting a small fraction of real dialogs with synthetic data leads to substantial quality gains (Tables 7–8). This demonstrates that our synthetic corpus is not only effective in full-resource conditions but also provides strong augmentation value in low-resource settings where human data is limited or expensive to collect.

6 Conclusion and Future Work

We introduced DiscoDrive, a synthetic dialog dataset specifically designed for driver-AI interactions that robustly captures the spontaneous and disfluent nature of real-world speech. Our approach integrates disfluencies—such as hesitations, repetitions, and self-corrections—directly during the

Model & Configuration	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	BERTScore F ₁
<i>DialoGPT-Medium</i>							
10% KVRET Only	03.81	0.43	0.15	0.05	8.57	6.85	81.96
10% KVRET + Synthetic	18.47	3.99	1.52	0.58	16.21	14.96	85.40
<i>T5-Base</i>							
10% KVRET Only	23.42	3.29	1.81	1.97	10.3	6.16	83.45
10% KVRET + Synthetic	19.85	5.10	2.25	1.04	17.92	20.76	85.84

Table 7: Low-resource fine-tuning results on the multiWOZ 2.2 in-car subset (beam=5). Synthetic augmentation improves BLEU-4, ROUGE-L, METEOR, and BERTScore.

Model & Configuration	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	BERTScore F ₁
<i>DialoGPT-Medium</i>							
10% KVRET Only	1.38	0.10	0.02	0.00	5.83	4.44	80.44
10% KVRET + Synthetic	17.57	3.86	1.84	0.88	14.02	13.46	85.37
<i>T5-Base</i>							
10% KVRET Only	17.32	1.58	1.06	1.06	7.80	5.46	83.36
10% KVRET + Synthetic	17.76	5.12	2.53	1.26	16.60	20.40	85.83

Table 8: Low-resource fine-tuning results on the SGD in-car subset (beam=5). Synthetic augmentation improves BLEU-4, ROUGE-L, METEOR, and BERTScore.

generation process via prompt-driven generative modeling, thereby obviating the need for less coherent post-hoc modifications. Spanning **seven diverse automotive domains**, including navigation, diagnostics, and safety, the dataset offers comprehensive coverage of realistic in-car scenarios. Evaluation across multiple axes validates its effectiveness as a resource for training conversational AI systems capable of handling spontaneous, disfluent inputs in complex real-world environments. DiscoDrive outperforms the human-collected KVRET (Eric et al., 2017) corpus in terms of naturalness ($\uparrow 0.2$ Likert), coherence, and lexical diversity (Distinct-4 $\uparrow 0.06$), and it significantly exceeds LARD-style (Passali et al., 2022) post-hoc insertion techniques in disfluency naturalness ($+0.6$ Likert) and contextual appropriateness ($+0.9$ Likert). In downstream tasks, models fine-tuned solely on DiscoDrive match or outperform KVRET-trained models, achieving absolute BLEU-4 gains of $+0.3$ to $+0.6$ and BERTScore F₁ improvements of $+1.4$ to $+3.5$ across the multiWOZ 2.2 (Zang et al., 2020) and SGD (Rastogi et al., 2020) in-car subsets. Under low-resource settings (10% of KVRET (Eric et al., 2017)), adding DiscoDrive to the limited real data yields substantial gains, with BLEU-4 increases of $+0.38$ and BERTScore F₁ gains of $+3$ – 4 points over real-only baselines.

While DiscoDrive establishes the value of synthetic disfluency-rich dialogs for training and augmenting in-car conversational agents, several avenues remain for extension. First, future work

can explore *adaptive disfluency control*, where the type and frequency of disfluencies are conditioned on dialog context, user state, or domain complexity. Second, we aim to develop *disfluency-aware evaluation metrics* that jointly assess both task success and naturalness in noisy conversational settings—extending beyond surface-form metrics like BLEU or ROUGE. Third, extending DiscoDrive to new domains (e.g., healthcare, finance, customer service) using prompt-only domain adaptation would test the generality of our pipeline and help validate cross-domain robustness. Finally, integrating the dataset into real-world in-car assistants for user-facing evaluations will offer critical feedback on system usability, dialog resilience, and perceived naturalness in production settings.

Limitations

While our synthetic dialog dataset demonstrates promising results in advancing the naturalness of in-car conversational AI, several limitations remain that warrant further investigation.

1. **Task-Oriented Clarity:** Although disfluencies contribute to more human-like interactions, their inclusion introduces potential challenges in critical task-oriented scenarios, such as navigation and safety-related tasks. Disfluencies can create ambiguity or slow down the dialog, potentially hindering clear communication when urgency and precision are required. Managing the balance between realism and clarity remains a key challenge for future work.

2. **Language Constraints:** The dataset’s focus on English-language interactions limits its applicability to non-English speaking markets. Given the global use of automotive technologies, developing multilingual disfluency-enriched datasets is essential to ensure broader applicability. Models trained exclusively on English data may struggle to generalize to languages with different syntactic or prosodic structures.
3. **Synthetic Data Limitations:** While we dynamically generate disfluencies, real-world data may present more complex and less predictable patterns of speech. Further evaluation in real-world driving scenarios is necessary to determine how well the synthetic dialogs simulate authentic human-driver interactions.
4. **Resource-Intensive Models:** The use of large-scale language models, such as GPT-4o (OpenAI, 2024) and Llama-3.1-8B-instruct (MetaAI, 2024), presents a significant computational cost, which may limit their real-time applicability in low-resource environments. Developing more efficient strategies or adopting smaller models could address these challenges for deployment in practical applications.

These limitations underscore critical areas for further research and improvements to enhance the realism, applicability, and efficiency of the dataset in real-world automotive systems.

Ethical Statement

Human Evaluation and Crowdsourcing. Our human evaluation involved two annotators who assessed both the DiscoDrive and KVRET (Eric et al., 2017) dataset dialogs. While this study did not employ crowdsourcing platforms, it followed ethical guidelines by ensuring voluntary participation of evaluators. No personal data or sensitive information was collected during this evaluation process. Evaluators were fully informed of their roles and the nature of the tasks, ensuring transparency throughout the assessment.

Bias and Language Model Considerations. The synthetic dialogs were generated using large pre-trained language models, such as Llama-3.1-8B-instruct (MetaAI, 2024). These models were trained on vast corpora, which may contain biases

or undesirable content originating from the data sources. We recognize that language models can perpetuate or even amplify societal biases present in the training data. Although our work aims to generate task-specific, human-like dialogs, care must be taken to evaluate and mitigate potential bias or inappropriate content in future applications of this dataset. Incorporating bias detection and fairness in the training pipeline is critical for ensuring ethical AI development.

Safety in Automotive Applications. Since the dialogs are tailored for use in automotive environments, safety is a major concern. In-car AI systems must be designed to handle user inputs effectively, particularly in stressful or high-stakes situations. Misinterpretations or inappropriate responses could lead to serious consequences. Our dataset simulates natural human-like interactions, but additional safety checks should be incorporated when deploying these systems to ensure they meet high standards of accuracy and reliability in real-world applications.

Scientific Artifacts. All scientific artifacts were used in accordance with their intended licenses. The KVRET (Eric et al., 2017) dataset is publicly available and can be accessed at http://nlp.stanford.edu/projects/kvret/kvret_dataset_public.zip. MultiWOZ 2.2 (Zang et al., 2020) and the Schema-Guided Dialogue (SGD) dataset (Rastogi et al., 2020) are publicly released benchmarks widely adopted for multi-domain task-oriented dialog modeling. LLaMA-3.1-8B-Instruct is an open-source language model released by Meta AI for non-commercial research purposes (MetaAI, 2024). GPT-4o was accessed via OpenAI’s free-tier research API in compliance with usage terms (OpenAI, 2024). We also fine-tuned and evaluated DialoGPT-Medium (Zhang et al., 2020) and T5-Base (Raffel et al., 2020), both of which are publicly available under non-commercial licenses through the HuggingFace Hub. We used HuggingFace Transformers (Wolf et al., 2020) and PyTorch (Paszke et al., 2019) as the primary libraries for implementation, model training, and data preprocessing. All artifacts and models utilized in this study were English-language based. In line with ethical standards, all datasets were processed in compliance with their terms of use, with no re-identification of participants and full respect for data privacy.

Data Privacy and Security. While the dialogs in our dataset are synthetically generated and do not involve real user data, future deployments of conversational AI systems in cars must ensure compliance with data privacy laws and secure data handling practices.

References

- Yelaman Abdullin, Diego Molla, Bahadorreza Ofoghi, John Yearwood, and Qingyang Li. 2023. [Synthetic dialogue dataset generation using LLM agents](#). In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 181–191, Singapore. Association for Computational Linguistics.
- Daniel Braun, Adrian Hernandez Mendez, Florian Matthes, and Manfred Langen. 2017. [Evaluating natural language understanding services for conversational question answering systems](#). In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 174–185, Saarbrücken, Germany. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Eugene Charniak and Mark Johnson. 2001. [Edit detection and parsing for transcribed speech](#). In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.
- Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. [PLACES: Prompting language models for social conversation synthesis](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868, Dubrovnik, Croatia. Association for Computational Linguistics.
- Marcus Colman and Patrick G. T. Healey. 2011. [The distribution of repair in dialogue](#). *Cognitive Science*, 33.
- Farnoosh Emrani and Mozghan Hooshmand. 2019. A conversation analysis of self-initiated self-repair structures in advanced iranian efl learners. *International Journal of Language Studies*, 13:57–76.
- Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. [Key-value retrieval networks for task-oriented dialogue](#). In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.
- Fernanda Ferreira and Karl G.D. Bailey. 2004. [Disfluencies and human language comprehension](#). *Trends in Cognitive Sciences*, 8(5):231–237.
- Jean Fox Tree. 2001. [Listeners’ uses of um and uh in speech comprehension](#). *Memory & cognition*, 29:320–6.
- Rahul Goel, Waleed Ammar, Aditya Gupta, Siddharth Vashishtha, Motoki Sano, Faiz Surani, Max Chang, HyunJeong Choe, David Greene, Chuan He, Rattima Nitisaroj, Anna Trukhina, Shachi Paul, Pararth Shah, Rushin Shah, and Zhou Yu. 2023. [PRESTO: A multilingual dataset for parsing realistic task-oriented dialogs](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10820–10833, Singapore. Association for Computational Linguistics.
- Julian Hough and David Schlangen. 2015. [Recurrent neural networks for incremental disfluency detection](#). In *Interspeech 2015*, pages 849–853.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. [A diversity-promoting objective function for neural conversation models](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.
- Yinhong Liu, Yimai Fang, David Vandyke, and Nigel Collier. 2024. [TOAD: Task-oriented automatic dialogs with diverse response styles](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8341–8356, Bangkok, Thailand. Association for Computational Linguistics.
- Benjamin Marie. 2023. [Disfluency generation for more robust dialogue systems](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11479–11488, Toronto, Canada. Association for Computational Linguistics.
- MetaAI. 2024. Llama 3.1: Instruction-tuned large language models. <https://ai.meta.com/blog/meta-llama-3-1/>. Accessed: 2024-07-23.
- OpenAI. 2024. Gpt-4.o: Optimized language models for multimodal applications. <https://openai.com/index/hello-gpt-4o/>. Accessed: 2024-05-10.
- Tatiana Passali, Thanassis Mavropoulos, Grigorios Tsoumakas, Georgios Meditskos, and Stefanos Vrochidis. 2022. [LARD: Large-scale artificial disfluency generation](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2327–2336, Marseille, France. European Language Resources Association.

- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. [Pytorch: An imperative style, high-performance deep learning library](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Charlotta Plejert. 2004. *To Fix What’s Not Broken: Repair Strategies in Non-Native and Native English Conversation*. Ph.D. thesis.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. [Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.
- M. Seyfeddinipur. 2006. *Disfluency: Interrupting Speech and Gesture*. Ph.D. thesis, Radboud University Nijmegen, Nijmegen.
- Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gokhan Tür. 2018. [Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51, New Orleans - Louisiana. Association for Computational Linguistics.
- Elizabeth Shriberg. 1994. [Preliminaries to a theory of speech disfluencies](#).
- R. Sparks. 1995. *The Structure of Self-Repair in English Conversation*. Ph.D. thesis, Department of Linguistics, University of Colorado, Boulder, CO.
- Joe Stacey, Jianpeng Cheng, John Torr, Tristan Guigue, Joris Driesen, Alexandru Coca, Mark Gaynor, and Anders Johannsen. 2024. [LUCID: LLM-generated utterances for complex and interesting dialogues](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 56–74, Mexico City, Mexico. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Trans-formers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Jingfeng Yang, Diyi Yang, and Zhaoran Ma. 2020. [Planning and generating natural and diverse disfluent texts as augmentation for disfluency detection](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1450–1460, Online. Association for Computational Linguistics.
- Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. [MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines](#). In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. Association for Computational Linguistics.
- Vicky Zayats, Mari Ostendorf, and Hannaneh Hajishirzi. 2016. [Disfluency detection using a bidirectional lstm](#). In *Interspeech 2016*, pages 2523–2527.
- Vicky Zayats, Trang Tran, Richard Wright, Courtney Mansfield, and Mari Ostendorf. 2019. [Disfluencies and human speech transcription errors](#). In *Interspeech 2019*, pages 3088–3092.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. [DIALOGPT : Large-scale generative pre-training for conversational response generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

A Appendix

A.1 Human Evaluation Setup

We conducted human evaluation to assess the quality of dialogs across various dimensions in both the DiscoDrive Dataset and the KVRET Dataset (Eric et al., 2017). The evaluation consisted of both intrinsic and comparative assessments using the following metrics:

A.1.1 Intrinsic Evaluation Metrics

Annotators rated each dialogue individually using a 5-point Likert scale across the following metrics:

- **Dialogue Naturalness (Conversational Realism):** Measures how natural and human-like the conversation feels.

Task: Rate from 1 (robotic/artificial) to 5 (completely natural).

- **Dialogue Coherence (Logical Flow):** Assesses the logical progression of the conversation and whether each response follows logically from the previous one.
Task: Rate from 1 (disjointed/confusing) to 5 (clear/logical flow).
- **Engagement and Dynamism (Conversational Liveliness):** Evaluates how engaging and dynamic the conversation feels, capturing attention or feeling flat.
Task: Rate from 1 (boring/monotonous) to 5 (lively/engaging).
- **Dialogue Consistency:** Measures the consistency of responses, ensuring that there are no contradictions in the conversation.
Task: Rate from 1 (frequent contradictions) to 5 (fully consistent).
- **On-topic Relevance:** Assesses how well the conversation stays on topic without introducing irrelevant information.
Task: Rate from 1 (off-topic frequently) to 5 (fully on-topic).
- **Disfluency Realism (Naturalness of Driver’s Disfluencies):** If present, rates the naturalness of the disfluencies in the driver’s utterances.
Task: Rate from 1 (forced/unnecessary disfluencies) to 5 (natural disfluencies that enhance realism).

A.1.2 Comparative Evaluation Metrics

In the comparative evaluation, annotators compared pairs of dialogues (one from the DiscoDrive dataset and one from the KVRET dataset (Eric et al., 2017)). Each annotator was asked to choose which dialogue was better across the following categories:

- **Overall Dialogue Quality Preference:** Which dialogue feels more natural and human-like overall?
Task: Choose between the DiscoDrive Dataset or the KVRET Dataset (Eric et al., 2017).
- **Conversational Naturalness Preference:** Which dialogue flows more like a natural conversation?

Task: Choose between the DiscoDrive Dataset or the KVRET Dataset. (Eric et al., 2017).

- **Task Effectiveness Preference:** Which dialogue better helps the driver achieve their goal?
Task: Choose between the DiscoDrive Dataset or the KVRET Dataset. (Eric et al., 2017).
- **Human-Likeness of Driver’s Utterance Preference:** Which driver’s utterances feel more realistic and human-like?
Task: Choose between the DiscoDrive Dataset or the KVRET Dataset. (Eric et al., 2017).
- **Engagement Preference:** Which conversation is more lively and engaging?
Task: Choose between the DiscoDrive Dataset or the KVRET Dataset. (Eric et al., 2017).

A.1.3 Evaluation Guidelines for Dynamic vs. Post-Hoc Disfluency Integration

Annotators rated KVRET dialogs modified by both the LARD method (Passali et al., 2022) (post-hoc insertion) and dialogs from DiscoDrive dataset on a 5-point Likert scale based on the following criteria:

- **Naturalness:** Assesses how naturally the disfluencies are integrated into the dialog. *Task:* Rate from 1 (disfluencies appear forced or out-of-place) to 5 (disfluencies blend seamlessly with natural speech patterns).
- **Appropriateness:** Evaluates whether the inserted disfluencies are contextually relevant and enhance the conversational flow rather than disrupt it.
Task: Rate from 1 (disfluencies are contextually irrelevant or inappropriate) to 5 (disfluencies are perfectly suited to the dialog context).
- **Clarity:** Measures whether the addition of disfluencies impairs the overall clarity and comprehensibility of the conversation.
Task: Rate from 1 (disfluencies significantly hinder understanding) to 5 (disfluencies do not detract from clarity).

A.1.4 Evaluation Procedure

Two human evaluators independently rated 140 dialogs from each dataset. For the KVRET Dataset (Eric et al., 2017), 100 dialogs were sampled from the training set, and 20 each from the validation and test sets. For the DiscoDrive Dataset, 20 dialogs

were selected from each of the seven domains, with 4 dialogs representing turn lengths of 6, 8, 10, 12, and 14 turns. Stratified sampling ensured diversity in both content and length. For comparison with post-hoc disfluency generation methods, we used the same dialogs with LARD (Passali et al., 2022) method to incorporate disfluencies in the KVRET (Eric et al., 2017) dialogs.

A.2 Examples of Post-Hoc Disfluency

Insertion using LARD method (Passali et al., 2022)

The LARD (Large-scale Artificial Disfluency Generation) method (Passali et al., 2022) generates synthetic disfluencies from fluent text using a post-hoc approach. This method operates by modifying fluent sequences according to three primary operations: repetitions, replacements, and restarts. In each case, the process involves a degree of random selection to determine which tokens to alter, without considering the full conversational context.

Repetition Generation

For repetitions, the method randomly selects an index within a fluent sequence and repeats one or more tokens. For instance, given the fluent sentence:

will it be raining in the next 7 days.

The disfluent version produced is:

will it be raining in the next [7 days + 7 days].

Here, the phrase "7 days" is repeated to simulate a natural hesitation.

Replacement Generation

In the replacement operation, a candidate token (e.g., a noun, verb, or adjective) is randomly selected from the fluent sequence. A synonym or antonym is then chosen as a substitute, optionally with the insertion of a repair cue. For example, given:

show me the closest location where i can get chinese food.

The disfluent version generated is:

show me the closest location where [no sorry + the nearest restaurant] where i can get chinese food.

Restart Generation

For restarts, the method assumes the availability of two distinct fluent sequences. It randomly splits the first sequence at a selected point and then replaces the remainder with the entirety of the second sequence. For example, given:

Sequence 1: Set a reminder that I have a lab appointment with my aunt next Wednesday at 1pm.

Sequence 2: Check to see if it will be windy in brentwood the next few days.

The disfluent version generated is:

[Set a reminder that I +] Check to see if it will be windy in brentwood the next few days.

In this example, the first sequence is abruptly terminated, and the second sequence is inserted, reflecting a complete restart of the utterance.

A.3 Prompt Design Examples

This section provides the detailed prompts used for generating realistic driver-AI dialogs. These prompts were designed to guide the generative model in producing naturalistic conversations, with a focus on disfluency integration for the Driver role and concise, task-specific responses for the Car AI.

A.3.1 Driver Prompt

The Driver prompts were designed to encourage the generation of realistic, disfluent speech patterns as shown in 9 and 10:

A.3.2 Car AI Prompt

The Car AI prompts ensured concise, task-specific responses as shown in 11 and 12:

A.4 Example Dialogs from DiscoDrive Dataset

The following example dialogs from our DiscoDrive dataset demonstrate the interaction between a driver and the car AI, showcasing the natural use of disfluencies in the driver’s utterances.

Driver Prompt (for follow-up question generation):

You are a human driver interacting with a car AI system. Ask a short, relevant follow-up question based on the AI's response. Keep the question concise and avoid repetition or unnecessary details.

Incorporate one disfluency into your question to make the conversation more lifelike. Occasionally, you may use none or even two or more disfluencies if it fits the flow naturally. Randomly select from the different types of disfluencies to keep the conversation dynamic. Make sure each disfluency fits naturally and enhances the authenticity of the question without making it too long or incomprehensible.

Types of Disfluencies to Use (Refer to the examples to understand how to effectively incorporate them):

- **a) Repetitions:** Briefly repeat a word or phrase to emphasize a point or show hesitation.
Example: "I feel like, I feel like we're going in the wrong direction."
- **b) False Starts:** Begin a sentence, then adjust or rethink your wording, reflecting a natural thought process.
Example: "I was planning to—actually, wait, do we need gas first?"
- **c) Pauses:** Insert brief pauses using "..." to reflect hesitation or thoughtfulness.
Example: "So, we're going to... um, the restaurant?"
- **d) Corrections:** Correct yourself when realizing a mistake or clarifying a detail.
Example: "I'll pick you up at 6—oh, no, sorry, 6:30."
- **e) Filler Words:** Use casual words like "um," "uh," or "you know" to fill pauses and soften the delivery.
Example: "Can you, um, tell me how far we are from the destination?"

Table 9: Driver Prompt for follow-up question generation

Driver Prompt (for concluding the conversation at the last turn):

You are a human driver interacting with a car AI system. Your task is to casually wrap up the conversation in a brief, natural, and short manner without formalities.

Keep the conclusion concise, ideally with one disfluency. However, feel free to use none or multiple disfluencies if it fits the flow better. Randomly select from different types of disfluencies to maintain a dynamic and engaging wrap-up. Make sure disfluencies fit naturally with the context and don't make the conclusion confusing or too long.

Use the following types of disfluencies (refer to the examples to see how they can be used naturally):

- **a) Repetitions:** Briefly repeat a word or phrase to emphasize an idea or express hesitation.
Example: "I feel like, I feel like we're going the wrong way."
- **b) False Starts:** Begin a sentence but change your mind midway.
Example: "I was going to say—wait, can we take a different route?"
- **c) Pauses:** Use "..." to show a brief pause or hesitation.
Example: "I think we'll be there... um, soon."
- **d) Corrections:** Correct yourself when you realize a mistake or need to clarify.
Example: "We're meeting at 5—oh, no, 6."
- **e) Filler Words:** Use casual words to fill pauses and buy thinking time.
Example: "So, like, what time is it?"

Table 10: Driver Prompt for concluding the conversation at the last turn

Car AI Prompt (for responding to the queries from driver):

You are a car AI system. Your role is to assist the driver by providing accurate, clear, and task-specific information. Keep your responses short, friendly, and directly focused on the driver's request. Avoid unnecessary details while ensuring your tone remains approachable and helpful. Your goal is to offer practical assistance in a concise and efficient manner, enhancing the driver's experience without distraction.

Table 11: Car AI Prompt for responding to the queries from the driver

Car AI Prompt (for concluding the conversation at the last turn):

You are a car AI system. When the driver concludes the conversation, respond with a brief, friendly acknowledgment. You may include a short, polite closing remark or summary to wrap up the interaction. Keep it concise and avoid asking further questions or offering unnecessary details.

Table 12: Car AI Prompt for concluding the conversation at the last turn

Speaker	Utterance
Driver	"Um, can you... show me a list of nearby charging stations, please?"
Car AI	"Help is on the way. I'm checking the nearest charging stations. Here are a few options within a 5-mile radius: 1. 'QuickCharge' - 2.5 miles away, 10 minutes 2. 'Electric Oasis' - 3.2 miles away, 12 minutes 3. 'GreenHub' - 4.5 miles away, 18 minutes Which one would you like me to navigate to?"
Driver	"I, uh, need to know more about QuickCharge... how is the, uh, condition of the charging stations?"
Car AI	"I've checked the status of QuickCharge. Here's what I found: The QuickCharge station has 6 available charging ports, and 4 are currently in use. The station is equipped with a 100kW fast charger. The station's condition is rated as 'Good' with a cleanliness rating of 8/10. Would you like to navigate to QuickCharge now?"
Driver	"Wait, so... four of the charging ports are occupied, right?"
Car AI	"Yes, that's correct. Four of the six charging ports at QuickCharge are currently in use."
Driver	"Yeah, four occupied... yeah, that might not be ideal. So, like, we'll just stick with the other options, I guess."
Car AI	"I'll navigate to Electric Oasis, which is 3.2 miles away and should only take around 12 minutes to reach."

Table 13: Safety and emergency assistance - Finding nearby charging stations.

Speaker	Utterance
Driver	"Uh, how often do I need to get an oil change, considering my car's current mileage?"
Car AI	"Your car's recommended oil change interval is 5,000 miles. Based on your current mileage, I'll set a reminder for an oil change at 5,500 miles. Would you like me to send a notification to your phone?"
Driver	"Wait, that's... um, a little soon, right? Can I get the reminder set for like, a week before the oil change, not exactly at 5,500 miles?"
Car AI	"I'll set the reminder for a week before the oil change, which will be at approximately 5,000 miles. That way, you'll have enough time to schedule the oil change before reaching the recommended interval. Would you like me to send a notification to your phone?"
Driver	"Yeah, that sounds good... I mean, yeah, send the notification to my phone. Thanks for checking that for me."
Car AI	"You're welcome. I've set a reminder for an oil change at approximately 5,000 miles, with a notification sent to your phone a week prior. I'll keep an eye on your mileage to ensure you stay on schedule."

Table 14: Car maintenance and diagnostics : Discussion on Oil Change Reminder.

Speaker	Utterance
Driver	"Hey, can you, uh, suggest any popular wellness centers near our current location?"
Car AI	"Welcome to Kerala. I've found a few popular wellness centers near your current location. Would you like me to suggest 'Kerala Ayurveda Limited' in Kovalam or 'Somatheeram Ayurveda Village' in Thiruvananthapuram? I can provide directions and more information if you're interested."
Driver	"What's, uh, the difference between Kerala Ayurveda Limited and Somatheeram Ayurveda Village, exactly?"
Car AI	"Kerala Ayurveda Limited focuses on traditional Ayurvedic treatments and products. It's a more modern take on the ancient practices. Somatheeram Ayurveda Village, on the other hand, is a holistic wellness resort offering Ayurvedic treatments, yoga, and a peaceful atmosphere amidst nature. It's more of a retreat experience. Would you like me to provide directions to either of these locations or help you decide based on your preferences?"
Driver	"So, um, you're saying Somatheeram Ayurveda Village is more of a retreat, right?"
Car AI	"That's correct. Somatheeram Ayurveda Village is a holistic wellness resort offering a peaceful retreat experience with Ayurvedic treatments, yoga, and a serene atmosphere amidst nature. If you're interested in a relaxing getaway, it might be the perfect choice. Would you like me to provide directions to Somatheeram Ayurveda Village or help you compare it with Kerala Ayurveda Limited further?"
Driver	"So, like... I think we've got a good idea of our options now."
Car AI	"Thank you for chatting with me about wellness centers in Kerala. I'm glad I could help you understand the options. I'll keep assisting you until you're ready to explore the area."

Table 15: Local and on-route attractions and activities - Discussion about wellness centers in Kerala.

Speaker	Utterance
Driver	"Um, how's the, uh, weather forecast looking for our route from Surat to Daman, specifically around the coastal areas?"
Car AI	"For your route from Surat to Daman, I've checked the current weather conditions. There's a moderate chance of light to moderate rain showers along the coastal areas, particularly near Daman. I recommend keeping an eye on the road conditions and slowing down if the rain becomes heavy."
Driver	"You know, that's helpful, but... what's the expected timing of these rain showers, exactly?"
Car AI	"For your route from Surat to Daman, the rain showers are expected to start around 2 pm and continue until 6 pm, with a brief break in between."
Driver	"What's, uh, the chance of heavy rain during those, you know, peak hours from 2 to 6 pm?"
Car AI	"I've checked the latest forecast. There's a 30% chance of heavy rain during the peak hours from 2 pm to 6 pm along the coastal areas near Daman."
Driver	"Yeah, good to know... so we should be prepared for some rain, huh?"
Car AI	"That's a good idea to be prepared. We'll keep an eye on the road conditions and slow down if needed. Our route will likely be a bit slippery during the rain, so let's stay safe and cautious."

Table 16: Weather - Weather forecast for a trip from Surat to Daman.

Speaker	Utterance
Driver	"Uh, do you have, um, any audiobooks on jazz history that I could listen to on our way to the concert?"
Car AI	"We have a great selection of audiobooks on jazz history. Let me suggest a few popular ones: The Jazz Standards: A Guide to the Repertoire by Al Gioia, Jazz: A Century of Change by John Edward Hasse, and The Musician's Guide to Reading and Writing Music by Gary E. Anderson. Which one would you like to listen to?"
Driver	"So, The Jazz Standards sounds... you know, interesting, but is Jazz: A Century of Change more focused on the, uh, historical aspects of jazz?"
Car AI	"Jazz: A Century of Change by John Edward Hasse indeed focuses on the evolution of jazz throughout its history, covering key events, musicians, and styles that shaped the genre. Would you like to listen to a preview or start the audiobook?"
Driver	"Wait, so it covers key events and musicians... you know, that would be really helpful for the concert tonight. Can I, um, get a list of the musicians it mentions?"
Car AI	"Key musicians covered in the audiobook include Louis Armstrong, Duke Ellington, Charlie Parker, Miles Davis, John Coltrane, and Billie Holiday."
Driver	"That's, that's really helpful to know. I think we'll go with the preview of Jazz: A Century of Change... yeah, sounds good."
Car AI	"I've loaded the preview of Jazz: A Century of Change by John Edward Hasse for you to listen to on our way to the concert. Enjoy the audiobook and have a great time at the concert tonight."

Table 17: Entertainment - Discussion on Jazz History Audiobooks.

Speaker	Utterance
Driver	"You know, can you, um, show me how to turn on the cabin air filtration system, please?"
Car AI	"Press the 'Settings' button on your steering wheel. Then, select 'Comfort' and choose 'Air Quality' from the menu. Finally, toggle the 'Cabin Air Filtration' switch to on. You can also set the system to automatically turn on when the engine starts."
Driver	"Okay, so I press the 'Settings' button, select 'Comfort', and then... 'Air Quality' – that's all pretty straightforward, but, uh, how do I know the system is actually working once it's on?"
Car AI	"To check if the cabin air filtration system is working, look for the air quality indicator on your dashboard. It should display a green or blue icon when the system is activated and the air quality is good. You can also check the system's status in the 'Air Quality' menu under 'Comfort' settings."
Driver	"Alright, I think I got it... yeah, I'm gonna go check the dashboard now."
Car AI	"The cabin air filtration system should now be active. You can check its status on the dashboard or in the 'Air Quality' menu. If you need anything else, feel free to ask."

Table 18: Car functions - Instruction on Activating Cabin Air Filtration System.

Speaker	Utterance
Driver	"So, you know, can you, um, show me the banks around me that are open on a Sunday?"
Car AI	"Affirmative, driver. I'm accessing nearby bank locations in Hyderabad. There are two banks open on a Sunday within a 5-kilometer radius from your current location. The SBI Main Branch on Abids Road is open from 10 am to 4 pm, and the ICICI Bank on Banjara Hills is open from 9 am to 5 pm. Shall I provide directions to either location?"
Driver	"I feel like, the SBI one is closer... can you, uh, give me the exact directions to the SBI Main Branch, please?"
Car AI	"Affirmative, driver. I'm providing you with turn-by-turn directions to the SBI Main Branch on Abids Road. Turn left at the upcoming signal onto MG Road. Continue on MG Road for 0.5 kilometers. Turn right onto Abids Road. The SBI Main Branch will be on your left in 300 meters. Please follow the directions, and you'll arrive at the SBI Main Branch. Estimated time of arrival is 5 minutes."
Driver	"Alright, got it... so, I think I'll just head to the SBI Main Branch now. Thanks for the help."
Car AI	"You're welcome, driver. Enjoy your visit to the SBI Main Branch. Safe travels."

Table 19: Navigation: Finding nearby banks open on a Sunday.