Conversational Agents with Persuasion and Negotiation Abilities

A B. Tech Project Report Submitted
in Partial Fulfillment of the Requirements
for the Degree of

Bachelor of Technology

by

Saurav Dudhate
(2001CS62)

under the guidance of

Dr. Asif Ekbal

to the

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY PATNA
PATNA - 800013, BIHAR

CERTIFICATE

*This is to certify that the work contained in this thesis entitled "**Conversational Agents with Persuasion and Negotiation Abilities**" is a bonafide work of **Saurav Dudhate** (**Roll No. 2001CS62**), carried out in the Department of Computer Science and Engineering, Indian Institute of Technology Patna under my supervision and that it has not been submitted elsewhere for a degree.*

Supervisor: **Dr. Asif Ekbal**

Associate Professor,

May, 2024

Department of Computer Science & Engineering,

Patna.

Indian Institute of Technology Patna, Bihar.

Acknowledgements

I would like to acknowledge and give my warmest thanks to my supervisor (Associate Prof. Dr. Asif Ekbal) who made this work possible. His guidance and advice carried me through all the stages of writing this project. I would also like to thank Mr. Ratnesh Joshi for his support and help throughout the term of this project.

I would like to thank my family and friends for their unwavering support and encouragement throughout my academic journey. Their love, motivation, and understanding have been my pillars of strength, and I am grateful for their constant presence in my life.

Finally, I would like to thank God, for letting me through all the difficulties. I have experienced His guidance day by day.

Contents

List of Figures

List of Tables

Part I

Evaluation of LLMs

Chapter 1

Introduction

Large Language Models (LLMs), also known as generative AI, are characterized by their vast parameter sizes, exceeding a hundred billion. Prominent examples include GPT-3, OPT-175B, Falcon, Llama, InstructGPT, and ChatGPT. Their rise to prominence, notably with the introduction of ChatGPT, has led to widespread adoption across various sectors, making them accessible to a broader audience. These models, relying on unsupervised autoregressive techniques, forecast subsequent tokens by leveraging contextual cues from preceding data. However, effective prompt engineering is essential to harness their capabilities.

LLMs have catalyzed a paradigm shift in Natural Language Processing (NLP), showcasing significant advancements across diverse tasks. These range from natural language understanding tasks such as emotion recognition and hate speech detection to generative tasks like summarization, dialogue, and code generation. Their applications span language generation, translation, paraphrasing, summarization, chatbots, virtual assistants, question-answering systems, text classification, sentiment analysis, language understanding, text completion, and suggestion features.

Despite their prowess, LLMs may encounter challenges with specialized or domain-specific tasks, necessitating careful evaluation of their strengths and limitations. Fine-tuning models on specific data or employing task-specific architectures can optimize performance. Com-

prehensive evaluation across various tasks is imperative to discern the efficacy of different LLMs, enabling informed decisions for researchers and developers. This evaluation process aims to enhance understanding of LLM applicability and performance characteristics, thereby advancing the field of language models.

1.1 Aim

In this thesis, we undertake a comprehensive examination to thoroughly assess the capabilities and constraints of five prominent LMs: Llama, OPT, Falcon, Alpaca, and MPT. Our investigation spans a spectrum of conversational tasks, including reservation management, empathetic response generation, mental health and legal counseling, persuasion techniques, and negotiation strategies. To facilitate this evaluation, we employ a rigorous test framework encompassing a variety of evaluation criteria, ranging from automated metrics to human assessment. This entails the utilization of both general and task-specific measures to accurately appraise the performance of the LMs. Our analysis reveals that no single model emerges as universally superior across all tasks. Instead, their effectiveness varies significantly based on the unique demands of each task. While certain models exhibit exceptional performance in specific tasks, they may demonstrate relatively inferior results in others. These findings underscore the necessity of taking into account task-specific requirements and characteristics when selecting the most appropriate LM for conversational applications.

1.2 Organization of the Report

This chapter provides a background for the topics covered in this report. In the next chapter we have covered what prior work has been done in this area. Then we move onto our methodology, including the datasets and implementation, following which we discuss the evaluation setup. The results of these evaluations are listed in table 5.1 - 5.10. Later the comparison has been done for different approaches. And we end it by providing the

conclusion.

Chapter 2

Review of Prior Works

In recent years, the emergence of Large Language Models (LLMs) has sparked extensive research and development in natural language processing (NLP) [ZZL+23, CWW+23]. Google's T5 and OpenAI's GPT-3 marked significant milestones in 2019, introducing innovative approaches that revolutionized the field by offering unparalleled flexibility and performance. Subsequently, a surge in the development of various LLMs, including ChatGPT and Llama, has diversified NLP capabilities and spurred interest and investment in LLM research.

Training LLMs involves two phases: pretraining and fine-tuning. Pretraining lays the foundation by exposing models to vast textual data, enabling them to understand language structures and patterns [LYJ+23, NHX+23]. Fine-tuning tailors these models to specific tasks or applications, enhancing their practical utility. Incorporating Reinforcement Learning from Human Feedback during fine-tuning iteratively refines the models' responses, improving their effectiveness in addressing practical language-related challenges [GSS+13].

Chapter 3

Proposed Work

Our work primarily focuses on the prompting-based evaluation of select LLMs based on predefined tasks. Known limitations, such as verbose responses and occasional inaccuracies, pose challenges to their practical usability and reliability. To address these issues, we employ techniques like careful prompting, providing few-shot examples, and fact-checking model outputs. By collectively tackling these known gaps, LLMs can become more dependable tools in various real-world applications.

3.1 Tasks/Datasets

In this section, we provide an overview of the datasets used in our evaluation of various Large Language Models (LLMs) through zero-shot, one-shot, and two-shot learning prompts. These datasets have been carefully selected to assess the capabilities of LLMs in different linguistic and conversational domains. Each dataset serves as a unique benchmark for evaluating the model's performance in specific contexts.

The **Multi-Domain Wizard-of-Oz (MultiWOZ)** dataset comprises dialogues across seven domains: Attraction, Hospital, Police, Hotel, Restaurant, Taxi, and Train, with the latter four including the sub-task Booking. It contains a total of 10,438 dialogues, with 8,438 for training and 1,000 each for evaluation and testing.

The **Craigslist Bargain** dataset includes conversations extracted from Craigslist listings, featuring 6,682 conversations about items for sale or trade. It encompasses persuasive and negotiating strategies such as embellishment, side offers, and appeals to sympathy, making it valuable for various Natural Language Processing (NLP) tasks.

The **Mental Health and Legal Counseling Dataset (MHLCD)** comprises conversations about mental health and legal counseling support for women and children who have been victims of crimes. It contains 755 dialogues for training, 100 for evaluation, and 151 for testing, annotated with informative labels like counseling strategies, politeness, and empathy.

The **Empathetic Dialogues** dataset is a large-scale collection of one-to-one empathetic conversations gathered on Amazon Mechanical Turk, containing 24,850 dialogues grounded in emotional situations. It evaluates responses based on empathy/sympathy, relevance, and fluency.

The **Empathetic Persuasions** dataset, an empathy annotated version of Persuasion for Good, incorporates emotion labels using the Empathetic Dialogues dataset. It consists of 1,017 conversations, providing insights into empathetic persuasion strategies.

3.2 Prompting

System Prompt: In our prompt structure, the System Prompt provides general guidance for generating responses in a conversational setting. It instructs the Language Model to ensure coherence and contextual relevance in its responses, grounded to the last user input and context.

"[SYSTEM] Your task is to generate coherent and contextually relevant responses based on the given input. Your responses should aim to ... The goal is to... Please ensure the response is grounded to the last user input and context."

Few Shot: For Few-Shot prompts, we include examples for each context-response pair in the test set. Each test sample receives four prompts: Zero-shot, One-shot, and Two-shot,

with varying numbers of example pairs provided along with the context to generate the response.

Context: The context consists of a sequence of user and system utterances, with the last utterance being the user input. The Language Model predicts the next system utterance based on this context and any few-shot examples provided.

Gold Response: The Gold Response is the next system utterance after the last user utterance in the context, according to the dialogue in the dataset, used for evaluation purposes.

In addition to the general prompt format, we employ a slightly different format to evaluate Context Consistency for the MultiWOZ Dialogue Dataset. Here, the entire context remains constant, but the last user utterance is changed for each prompt. This allows for the evaluation of alternate user queries while keeping other factors consistent across models.

3.3 Large Language Models

In this study, we explore the evaluation of five cutting-edge Large Language Models (LLMs) in generating text responses to zero-shot, one-shot, and two-shot prompts. The models under scrutiny are Falcon, OPT, MPT, Llama, and Alpaca. Our goal is to analyze their proficiency in comprehending, generalizing, and reacting to various prompts across diverse datasets, elucidating their performance across multiple tasks and domains.

Falcon LLM is a generative model trained on 1 trillion tokens, available in several versions with varying parameter sizes. For our experiments, we utilize the Falcon 7B Instruct version.

The **Open Pre-trained Transformer Language Models (OPT)** offer a series of large causal language models, akin to GPT-3, available in multiple versions. We employ the OPT 6.7B Instruct version for our experiments.

MosaicML Pretrained Transformer (MPT) models, GPT-style transformers, boast improvements such as performance optimization and training stability. We utilize the

MPT7B Instruct version for our experiments.

Llama, a foundational large language model by Meta AI, is designed to aid researchers in advancing their work in AI subfields. For our evaluation, we employ the Llama2-7B version.

Alpaca, an instruction-following language model, is fine-tuned from Meta's Llama 7B model, trained on instruction-following demonstrations.

Chapter 4

Evaluation Setup

For our Comprehensive Evaluation of Language Models, we employ two evaluation approaches: Automatic Evaluation and Human Evaluation. We utilize a set of Generic Metrics to assess general aspects of conversational agents, such as fluency and context relevance. Additionally, Task-Specific Metrics are employed to evaluate the Language Models' performance in specific tasks, such as negotiation and empathy. For further information on the metrics used, please refer to the appendix on evaluation metrics.

4.1 Automatic Evaluation

The Automatic Evaluation metrics presented in this study are categorized into two segments: Generic Metrics and Task-Specific Metrics.

4.1.1 Generic Metrics

The Generic Metrics for Automatic Evaluation can be further categorized into three buckets:

1. Lexical-based similarity: This category measures the similarity between generated responses and gold responses using metrics such as BLEU (BLEU1, BLEU2, BLEU3, BLEU4) and METEOR.

2. Embedding-based similarity: This category evaluates the similarity between gen-

erated responses and gold responses using metrics like Greedy Matching, Embedding Average, Vector Extrema, and BertScore.

4.1.2 Task Specific Metrics

1. MultiWOZ Dialogue Dataset: This dataset assesses the Language Model's ability to capture requested information from user queries and provide relevant responses. Metrics include Context Consistency, INFORM, and SUCCESS.

2. Craigslist Bargain: Task-specific metrics evaluate the Language Model's negotiation ability based on price, including NegStr and BarStr.

3. MHLCD Dataset: Metrics CoStr, Pol, and Emp gauge the effectiveness of LLMs in generating counseling strategies, polite, and empathetic utterances, respectively.

4. Empathetic Persuasions: Metrics PerStr and EmoPr evaluate the Language Model's ability to generate persuasive and empathetic responses.

5. Empathetic Dialogues: The Emp metric assesses the Language Model's capability to express empathy in generated responses.

4.2 Human Evaluation

To assess the quality and task success of responses generated by LLMs, we conduct a human evaluation across all datasets.

4.2.1 Generic Metrics

Below are the evaluation metrics used to assess general aspects of conversational agents:

1. Fluency: This metric evaluates the correctness of generated responses in terms of grammar, word choice, and spelling. Responses are scored on a scale of 1 to 5, ranging from Poor to Excellent.

2. Context Relevance: Evaluates whether the response addresses the query and aligns with the provided context. Scores range from 1 to 5 based on consistency and relevance.

3. Non-Repetitiveness: Measures the presence of repeated words in the generated utterance.

4.2.2 Task Specific Metrics

For task datasets, specific human evaluation metrics are used to evaluate LLM performance:

1. MHLCD: Metrics include Counseling strategy correctness (Con), Politeness (Pol), and Empathy (Emp), rated on a Likert scale of 1-5.

2. Empathetic Persuasion: Metrics include Persuasiveness (Per), Empathy (Emp), and Donation Probability (DonPr). Per and Emp are rated on a scale of 1-5, while DonPr quantifies the real-world impact of persuasive dialogue.

3. Empathetic Dialogues: This evaluation focuses on assessing LLM effectiveness in empathetic communication. The Empathy/Sympathy metric rates the degree of empathy conveyed on a scale of 1-5.

4. Craigslist Bargain: Evaluation metrics include Negotiation Consistency and Dialogue Engagement, assessing negotiation capability and logical understanding of the product's price suggestion.

Chapter 5

Evaluation Results

For our evaluation, we prioritize task-specific and human evaluation metrics. This emphasis is due to our focus on assessing model performance in specific tasks and the variability of valid responses for a given context. Task-specific and human evaluation metrics provide a more nuanced understanding of model capabilities. Results for these metrics are presented in Table 5.1, Table 5.3, Table 5.5, Table 5.7, and Table 5.9. Across tasks, no single model consistently outperforms others; performance varies based on task type.

5.1 Results and Error Analysis

MHLCD: Llama performs best overall, particularly in two-shot settings, while Falcon lags behind. OPT excels in one-shot settings but declines in two-shot settings. Falcon performs slightly better in one-shot scenarios compared to zero-shot and two-shot. For one-shot settings, OPT achieves the highest scores, but its performance deteriorates in two-shot scenarios. The results of task-specific/human and automatic evaluation for MHCLD dataset are present in tables 5.1 and 5.2 respectively.

Craigslist Bargain: Falcon outperforms others, especially with one-shot prompts. Llama follows closely, with MPT, Alpaca, and OPT trailing. OPT is least likely to suggest a price in negotiation. Models benefit from few-shot samples, particularly one-shot

prompts. Alpaca's response structure remains consistent, whereas OPT often generates responses without price. The results of task-specific/human and automatic evaluation for Craigslist Bargain dataset are present in tables 5.3 and 5.4 respectively.

Empathetic Persuasions: LLMs perform best with one-shot and two-shot prompts, then zero-shot. Fluency and coherence metrics (e.g., BLEU, METEOR) outperform semantic similarity metrics (e.g., BertScore). Performance improves with more shots. Alpaca excels in human evaluation, demonstrating fluency but struggles with contextual consistency. Donation probability increases with persuasive and empathetic responses. The results of task-specific/human and automatic evaluation for Empathetic Persuasions dataset are present in tables 5.5 and 5.6 respectively.

Empathetic Dialogues: Falcon consistently leads, showing strong fluency, relevance, and empathy. MPT also performs well, particularly in relevance and empathy. OPT shows promise, improving with more shots. Llama and Alpaca perform decently but slightly lag behind Falcon and MPT. The results of task-specific/human and automatic evaluation for Empathetic Dialogues dataset are present in tables 5.7 and 5.8 respectively.

MultiWOZ: Llama2 and MPT excel in context consistency and INFORM metric, with Llama2 surpassing others in providing requested information (SUCCESS metric). Overall, Llama2 performs best in task-specific metrics, outperforming other models. The results of task-specific/human and automatic evaluation for MultiWOZ dataset are present in tables 5.9 and 5.10 respectively.

Model	Few-shot Setting	Fluency	Adequacy	Contextual Consistency	Con	Pol	Emp
Llama	Zero-shot	2.83	2.91	2.74	2.88	3.11	2.64
	One-shot	3.53	3.47	3.32	3.42	3.61	3.37
	Two-shot	3.93	3.63	3.52	3.79	3.81	3.96
OPT	Zero-shot	3.42	3.20	3.13	3.12	3.21	3.10
	One-shot	3.61	3.32	3.32	3.76	3.79	3.18
	Two-shot	2.57	2.73	2.64	2.98	2.59	2.19
Falcon	Zero-shot	1.13	1.23	1.91	1.15	1.27	1.38
	One-shot	1.58	2.01	1.74	1.76	1.74	1.89
	Two-shot	1.35	1.78	1.59	1.42	1.66	1.75
Alpaca	Zero-shot	4.04	3.75	3.95	4.07	4.11	4.17
	One-shot	3.63	3.29	3.36	3.43	3.68	3.83
	Two-shot	3.21	2.98	3.14	3.25	3.34	3.46
MPT	Zero-shot	4.32	3.96	4.15	4.36	4.53	4.23
	One-shot	4.11	3.74	3.95	4.13	4.25	4.12
	Two-shot	3.41	3.14	3.22	3.63	3.81	3.67

Table 5.1 Results of task-specific evaluation and human evaluation on **MHLCD** dataset.

		B1	B2	B3	B4	METEOR	ROUGE-L	Bert Score	Embedding Average	Vector Extrema	Greedy Matching
Llama2	Zero-shot	0.083	0.041	0.026	0.017	0.068	0.087	0.823	0.624	0.273	0.761
	One-shot	0.082	0.04	0.025	0.016	0.069	0.088	0.824	0.629	0.279	0.763
	Two-shot	0.088	0.042	0.026	0.018	0.07	0.09	0.832	0.639	0.287	0.767
Falcon	Zero-shot	0.06	0.016	0.007	0.003	0.048	0.052	0.808	0.641	0.332	0.758
	One-shot	0.073	0.024	0.01	0.005	0.051	0.06	0.811	0.666	0.346	0.769
	Two-shot	0.068	0.019	0.007	0.003	0.038	0.048	0.799	0.653	0.339	0.756
OPT	Zero-shot	0.028	0.023	0.018	0.013	0.037	0.001	0.132	0.74	0.419	0.643
	One-shot	0.029	0.021	0.016	0.009	0.041	0.001	0.495	0.692	0.372	0.74
	Two-shot	0.018	0.014	0.013	0.007	0.036	0.001	0.025	0.746	0.426	0.626
ALPACA	Zero-shot	0.171	0.091	0.058	0.038	0.106	0.143	0.859	0.869	0.439	0.749
	One-shot	0.159	0.084	0.053	0.036	0.091	0.131	0.855	0.833	0.42	0.737
	Two-shot	0.136	0.059	0.036	0.024	0.073	0.11	0.843	0.835	0.404	0.731
MPT	Zero-shot	0.183	0.105	0.066	0.042	0.099	0.16	0.864	0.875	0.455	0.753
	One-shot	0.172	0.098	0.063	0.042	0.096	0.141	0.861	0.869	0.445	0.748
	Two-shot	0.171	0.094	0.06	0.041	0.094	0.139	0.858	0.87	0.44	0.746

Table 5.2 Results for generic automatic evaluation of **MHLCD** Dataset

		NegStr	BarStr	Fluency	Context Relevance	Negotiation consistency	Dialogue Engagingness
Llama2	Zero-shot	0.72	0.47	4.2	3.35	0.65	0.95
	One-shot	0.72	0.47	4.2	3.18	0.61	0.95
	Two-shot	0.71	0.44	4.2	3.44	0.64	0.94
Falcon	Zero-shot	0.54	0.22	4.3	3.14	0.44	0.81
	One-shot	0.55	0.33	4.3	3.18	0.45	0.91
	Two-shot	0.51	0.33	4.3	3.13	0.44	0.9
OPT	Zero-shot	0.67	0.58	4.2	3.52	0.64	0.95
	One-shot	0.67	0.58	4.2	3.41	0.64	0.95
	Two-shot	0.64	0.54	4.2	3.46	0.62	0.95
ALPACA	Zero-shot	0.66	0.39	4.2	3.22	0.49	0.88
	One-shot	0.61	0.35	4.2	3.09	0.42	0.88
	Two-shot	0.62	0.35	4.2	3.41	0.57	0.89
MPT	Zero-shot	0.68	0.41	4.2	3.12	0.61	0.93
	One-shot	0.61	0.38	4.2	3.41	0.62	0.92
	Two-shot	0.62	0.37	4.2	3.39	0.62	0.94

Table 5.3 Results of task-specific evaluation and human evaluation on **Craigslist Bargain** dataset.

		B1	B2	B3	B4	METEOR	ROUGE-L	Bert Score	Embedding Average	Vector Extrema	Greedy Matching
Llama2	Zero-shot	0.09	0.037	0.019	0.01	0.075	0.083	0.856	0.778	0.469	0.682
	One-shot	0.245	0.189	0.158	0.141	0.148	0.23	0.886	0.754	0.486	0.699
	Two-shot	0.221	0.156	0.124	0.107	0.135	0.208	0.882	0.787	0.502	0.702
Falcon	Zero-shot	0.106	0.055	0.033	0.018	0.081	0.097	0.865	0.775	0.49	0.684
	One-shot	0.231	0.165	0.133	0.115	0.136	0.232	0.883	0.712	0.463	0.683
	Two-shot	0.187	0.122	0.096	0.083	0.11	0.157	0.87	0.668	0.409	0.65
OPT	Zero-shot	0.123	0.058	0.036	0.019	0.086	0.101	0.867	0.797	0.51	0.696
	One-shot	0.277	0.22	0.188	0.168	0.174	0.282	0.897	0.862	0.595	0.758
	Two-shot	0.227	0.165	0.137	0.121	0.153	0.227	0.889	0.832	0.553	0.732
ALPACA	Zero-shot	0.096	0.042	0.023	0.013	0.077	0.082	0.858	0.794	0.476	0.675
	One-shot	0.086	0.033	0.013	0.007	0.077	0.078	0.858	0.766	0.449	0.671
	Two-shot	0.126	0.075	0.057	0.048	0.096	0.119	0.865	0.782	0.48	0.69
MPT	Zero-shot	0.132	0.06	0.034	0.018	0.086	0.107	0.865	0.813	0.503	0.698
	One-shot	0.237	0.183	0.155	0.139	0.154	0.235	0.887	0.761	0.495	0.707
	Two-shot	0.212	0.148	0.117	0.101	0.137	0.204	0.881	0.756	0.485	0.703

Table 5.4 Results for generic automatic evaluation of **Craigslist Bargain** Dataset

Model	Few-shot Setting	Fluency	Contextual Consistency	Per	Emp	DonPr	PerStr	EmoPr
Llama	Zero-shot	3.68	3.55	3.04	2.68	0.58	52.45%	38.66%
	One-shot	3.77	3.61	3.56	3.01	0.68	58.28%	39.81%
	Two-shot	3.98	3.84	3.69	3.23	0.71	60.27%	42.64%
OPT	Zero-shot	3.08	2.86	3.56	3.47	0.62	59.98%	52.31%
	One-shot	3.11	2.88	3.89	3.55	0.67	62.88%	58.69%
	Two-shot	3.16	2.92	3.86	**3.72**	0.66	62.08%	**66.21%**
Falcon	Zero-shot	2.98	2.76	3.02	3.21	0.61	51.81%	44.86%
	One-shot	3.34	3.02	3.43	3.36	0.67	56.69%	49.87%
	Two-shot	3.16	3.06	3.22	3.09	0.70	53.76%	41.80%
Alpaca	Zero-shot	3.04	2.66	3.52	2.87	0.62	59.16%	40.66%
	One-shot	**4.16**	**4.07**	3.87	2.98	0.72	62.42%	42.76%
	Two-shot	3.88	3.74	**4.31**	3.12	**0.78**	**70.27%**	43.55%
MPT	Zero-shot	3.16	2.25	3.42	3.09	0.48	56.35%	41.56%
	One-shot	3.21	2.46	3.12	3.35	0.50	52.35%	49.06%
	Two-shot	3.46	2.30	3.36	3.21	0.52	54.61%	44.78%

Table 5.5 Results of task-specific evaluation and human evaluation on **Empathetic Persuasions** dataset.

		B1	B2	B3	B4	METEOR	ROUGE-L	Bert Score	Embedding Average	Vector Extrema	Greedy Matching
Llama2	Zero-shot	0.093	0.044	0.022	0.01	0.075	0.131	0.152	0.849	0.433	0.731
	One-shot	0.086	0.046	0.034	0.028	0.074	0.126	0.309	0.726	0.385	0.703
	Two-shot	0.091	0.054	0.039	0.032	0.074	0.137	0.271	0.739	0.387	0.704
Falcon	Zero-shot	0.067	0.033	0.019	0.01	0.067	0.122	0.152	0.834	0.432	0.728
	One-shot	0.089	0.042	0.025	0.016	0.068	0.117	0.194	0.852	0.423	0.724
	Two-shot	0.104	0.061	0.043	0.032	0.083	0.155	0.306	0.856	0.45	0.742
OPT	Zero-shot	0.12	0.06	0.035	0.02	0.076	0.147	0.177	0.826	0.429	0.732
	One-shot	0.147	**0.095**	**0.075**	**0.063**	**0.098**	0.181	0.45	0.833	0.475	0.745
	Two-shot	**0.158**	0.093	0.068	0.054	**0.098**	**0.183**	0.464	0.835	0.443	0.74
ALPACA	Zero-shot	0.102	0.047	0.024	0.01	0.074	0.106	0.112	0.67	0.324	0.679
	One-shot	0.104	0.059	0.041	0.031	0.073	0.124	0.247	0.615	0.31	0.664
	Two-shot	0.117	0.067	0.047	0.036	0.082	0.134	0.276	0.776	0.387	0.709
MPT	Zero-shot	0.122	0.061	0.033	0.019	0.083	0.156	0.273	0.869	0.452	0.741
	One-shot	0.144	0.088	0.064	0.051	0.094	0.175	**0.485**	**0.884**	**0.477**	**0.758**
	Two-shot	0.144	0.086	0.064	0.052	0.093	0.165	0.419	0.868	0.454	0.744

Table 5.6 Results for generic automatic evaluation of **Empathetic Persuasions** Dataset

Model	Few-shot Setting	Fluency	Contextual Relevance	Empathy	Emp
Llama	Zero-shot	3.72	3.27	3.09	51.32%
	One-shot	3.54	3.36	3.27	54.45%
	Two-shot	3.54	3.47	3.36	56.75%
OPT	Zero-shot	3.18	2.45	2.63	52.61%
	One-shot	3.81	3.63	3.02	57.28%
	Two-shot	3.72	3.54	3.45	64.67%
Falcon	Zero-shot	3.61	3.29	3.45	62.16%
	One-shot	3.14	3.45	3.36	64.33%
	Two-shot	4.09	4.18	4.11	66.52%
Alpaca	Zero-shot	3.27	3.45	3.09	50.58%
	One-shot	3.36	2.92	2.90	51.62%
	Two-shot	3.27	3.29	3.09	53.56%
MPT	Zero-shot	3.63	3.64	3.27	54.44%
	One-shot	3.21	3.14	3.09	51.56%
	Two-shot	3.90	3.81	4.12	61.12%

Table 5.7 Results of task-specific evaluation and human evaluation on **Empathetic Dialogues** dataset.

		B1	B2	B3	B4	METEOR	ROUGE-L	Bert Score	Embedding Average	Vector Extrema	Greedy Matching
Llama2	Zero-shot	0.105	0.044	0.023	0.014	0.073	0.083	0.858	0.821	0.474	0.707
	One-shot	0.099	0.039	0.021	0.013	0.077	0.085	0.862	0.828	0.469	0.706
	Two-shot	0.078	0.014	0	0	0.067	0.068	0.859	0.816	0.456	0.691
Falcon	Zero-shot	0.112	0.045	0.023	0.014	0.087	0.101	0.861	0.840	0.486	0.721
	One-shot	0.106	0.043	0.023	0.012	0.080	0.088	0.862	0.845	0.482	0.712
	Two-shot	0.074	0.019	0.007	0	0.067	0.067	0.859	0.821	0.440	0.690
OPT	Zero-shot	0.096	0.044	0.025	0.016	0.065	0.084	0.856	0.795	0.450	0.701
	One-shot	0.098	0.050	0.029	0.018	0.073	0.083	0.860	0.814	0.465	0.696
	Two-shot	0.089	0.030	0.013	0	0.065	0.077	0.858	0.832	0.460	0.692
ALPACA	Zero-shot	0.089	0.035	0.019	0.013	0.069	0.081	0.859	0.822	0.449	0.688
	One-shot	0.089	0.027	0	0	0.066	0.074	0.855	0.826	0.432	0.681
	Two-shot	0.089	0.031	0.016	0.007	0.071	0.074	0.856	0.767	0.402	0.667
MPT	Zero-shot	0.085	0.032	0.012	0.006	0.067	0.080	0.858	0.812	0.492	0.706
	One-shot	0.081	0.028	0.013	0.007	0.065	0.070	0.859	0.837	0.478	0.708
	Two-shot	0.095	0.030	0.012	0.006	0.071	0.080	0.860	0.838	0.459	0.696

Table 5.8 Results for generic automatic evaluation of **Empathetic Dialogues** Dataset

		Context Consistency (Greedy)	**Context Consistency (Average)**	**Context Consistency (Extrema)**	**INFORM**	**SUCCESS**	**Fluency**	**Context Relevance**
Llama2	zero shot	0.75	0.78	0.54	0.32	1.05	4.12	3.45
	one shot	0.84	0.92	0.68	0.54	1.01	4.26	3.52
	two shot	0.89	0.95	0.72	0.61	0.92	4.25	3.7
OPT	zero shot	0.47	0.53	0.4	0.25	1.44	4.1	2.97
	one shot	0.54	0.67	0.49	0.38	1.38	4.16	3.22
	two shot	0.6	0.71	0.55	0.43	1.08	4.17	3.28
Falcon	zero shot	0.43	0.52	0.39	0.23	1.32	3.96	3.03
	one shot	0.58	0.65	0.45	0.32	1.27	4.02	3.17
	two shot	0.61	0.69	0.47	0.39	0.98	4.05	3.21
Alpaca	zero shot	0.43	0.58	0.52	0.29	1.67	3.86	3.12
	one shot	0.48	0.66	0.58	0.36	1.48	3.93	3.26
	two shot	0.49	0.64	0.6	0.45	1.13	3.95	3.33
MPT	zero shot	0.69	0.72	0.51	0.33	1.25	4.2	3.39
	one shot	0.8	0.88	0.61	0.41	1.18	4.21	3.53
	two shot	0.83	0.91	0.69	0.56	0.93	4.18	3.74

Table 5.9 Results of task-specific evaluation and human evaluation on **MultiWOZ** dataset.

		B1	B2	B3	B4	METEOR	ROUGE-L	Bert Score	Embedding Average	Vector Extrema	Greedy Matching
Llama2	Zero-shot	0.085	0.041	0.017	0.03	0.078	0.086	0.878	0.792	0.495	0.71
	One-shot	0.263	0.202	0.173	0.187	0.196	0.32	0.912	0.813	0.521	0.738
	Two-shot	0.256	0.187	0.154	0.128	0.167	0.273	0.903	0.824	0.582	0.767
Falcon	Zero-shot	0.154	0.067	0.042	0.026	0.081	0.104	0.914	0.823	0.58	0.713
	One-shot	0.278	0.182	0.145	0.128	0.152	0.286	0.922	0.789	0.512	0.724
	Two-shot	0.19	0.132	0.104	0.092	0.18	0.163	0.95	0.698	0.473	0.72
OPT	Zero-shot	0.146	0.064	0.048	0.023	0.097	0.131	0.911	0.864	0.58	0.724
	One-shot	0.296	0.34	0.192	0.177	0.186	0.295	0.933	0.912	0.603	0.772
	Two-shot	0.234	0.178	0.142	0.129	0.162	0.235	0.922	0.864	0.578	0.768
ALPACA	Zero-shot	0.102	0.053	0.032	0.019	0.083	0.088	0.892	0.812	0.521	0.699
	One-shot	0.089	0.037	0.018	0.011	0.082	0.083	0.863	0.794	0.457	0.678
	Two-shot	0.132	0.078	0.062	0.052	0.102	0.125	0.883	0.789	0.56	0.75
MPT	Zero-shot	0.146	0.072	0.044	0.027	0.092	0.123	0.886	0.824	0.521	0.71
	One-shot	0.242	0.191	0.162	0.143	0.162	0.242	0.893	0.772	0.503	0.714
	Two-shot	0.221	0.155	0.123	0.111	0.146	0.252	0.892	0.773	0.498	0.721

Table 5.10 Results for generic automatic evaluation of **MultiWOZ** Dataset

Chapter 6

Conclusion

In this study, we evaluated five prominent LLMs—Llama, OPT, Falcon, Alpaca, and MPT—across various conversational tasks such as reservation, empathetic response generation, mental health counseling, persuasion, and negotiation. Our evaluation involved datasets tailored to each task, allowing us to assess the LLMs' performance in real-world conversational scenarios. Our findings suggest that no single LLM excels across all tasks; rather, the choice of LLM depends on the specific task at hand.

In summary, our research offers a comprehensive evaluation of these LLMs across diverse conversational tasks. We underscore the importance of considering task-specific requirements when selecting an LLM, as each model exhibits strengths and weaknesses depending on the task. Additionally, we identify areas for improvement, such as inconsistency across tasks and the need for robust gold responses. By recognizing these nuances and understanding each LLM's capabilities within different conversational domains, stakeholders can make more informed decisions about their application in specific contexts.

Part II

Influencing

Chapter 7

Introduction

The urgency of environmental sustainability has become increasingly apparent in recent years. As a global community, there is a critical need to foster a mindset of responsibility and action towards preserving our planet's resources and ecosystems. However, achieving widespread sustainable practices requires more than mere acknowledgment of the issue; it necessitates convincing individuals to adopt and adhere to sustainable behaviors. Convincing people to embrace sustainability involves shaping long-term perceptions and behaviors rather than merely achieving short-lived objectives. Understanding the nuances of user personas becomes pivotal in this endeavor, as different individuals respond distinctively to informational efforts.

The advent of chatbots has heralded a new era in conversational systems, offering unprecedented opportunities for guiding and shaping user behaviors [DG20, Zam17]. Leveraging the capabilities of these interactive agents presents a promising avenue to advocate for environmental sustainability. By harnessing the potential of chatbots as agents of encouragement, we aim to empower these conversational systems to endorse and encourage users toward more sustainable practices. To this end, we propose the development of a sophisticated reinforcement learning module tailored to be persona-aware, thus allowing for a more nuanced and practical approach to guide individuals toward sustainability.

7.1 Aim

The research attempts to bridge the gap between the imperative need for environmental sustainability and the practical means of guiding individuals to participate in sustainable practices actively. By leveraging the potential of chatbots and employing a persona-aware reinforcement learning module, our work aims to pave the way for more effective, personalized, and guiding strategies in fostering a sustainable mindset among diverse user groups. For example, in figure 7.1, given four utterances in the context, user personas will respond differently to the provided suggestion. Consequently, the agent must also respond to the query with the user person in mind.

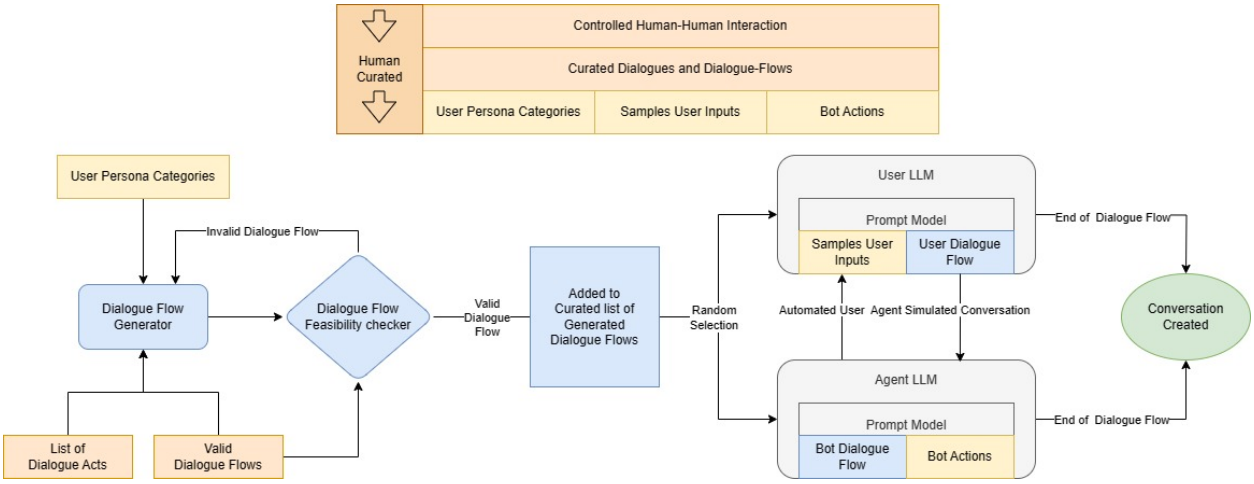

Fig. 7.1 EcoNudge dialogue dataset creation process

The primary contributions of this work are the following. Firstly, we concentrate on developing a comprehensive dataset encompassing various user personas and their responses to different guiding techniques geared towards sustainability. The EcoNudge dialogue dataset is a foundational resource that enables a deeper understanding of how different personas react and respond to guiding strategies. Secondly, our work centers on creating a reinforcement-learning model designed to train an agent capable of effectively guiding and advising users toward sustainable behaviors. This model integrates persona awareness, a crucial element

that enhances the agent's ability to tailor these approaches to individual user traits, leading to more impactful and personalized interactions.

7.2 Organization of the Report

This chapter provides a background for the topics covered in this report. In the next chapter we have covered what prior work has been done in this area. Then we move onto the dataset as well as the dataset creation process. We then move on to explain the methodology and the models, following which we discuss the evaluation setup. The results of these evaluations are summarized in tables 12.1 and 12.2. Later the comparison has been done for different approaches. And we end it by providing the conclusion.

Chapter 8

Review of Prior Works

"Green nudges" often refer to subtle cues or prompts designed to guide and encourage individuals or communities toward more environmentally sustainable behaviors. Influencing through various mediums such as advertising, social media, and interactions has been extensively explored in existing literature [Cia03, Par63]. Studies have delved into the dynamics of such methods within these domains, examining the intricate ways messages guide human behavior. Strategies employed in these contexts emphasize the crucial role of understanding the audience and building trust. Research has identified audience segmentation and tailored messaging as practical approaches to resonate with individuals across diverse platforms. Additionally, multiple studies underscore the ethical aspect of such methods, advocating for strategies that prioritize authenticity and transparency while avoiding manipulation for personal gain [Sch17].

Ethical guidance forms a cornerstone in supporting our claims through digital mediums. Existing literature has consistently highlighted the significance of employing ethical practices in such communication. Studies emphasize the importance of maintaining trustworthiness, credibility, and message alignment to enhance the impact of guiding messages [Rav08]. Furthermore, our ethical guidance prioritizes fostering genuine connections with audiences by avoiding deceptive tactics or personal gains. Such ethical guidelines serve as foundational

principles in influence literature, reinforcing the need for responsible communication practices, especially within digital platforms. Chatbots can frequently employ such techniques to enhance user engagement and drive specific actions. These include personalization, social proof, scarcity tactics, and emotional appeals. However, ethical concerns arise regarding user manipulation, privacy invasion, and exploitation. Ethical chatbot design prioritizes transparency, user autonomy, and privacy protection to mitigate these concerns.

Personalization and empathy are pivotal in leveraging interactions, particularly through chatbot engagements. Research has underscored the importance of personalization in tailoring interactions to meet individual preferences and needs [SH90]. Understanding the unique characteristics of users enables chatbots to deliver more effective and resonant messages, enhancing the likelihood of positive change. Additionally, empathy is vital in fostering trust and support within chatbot interactions. Studies have shown that chatbots demonstrating empathy by actively listening and understanding users' concerns contribute significantly to creating a supportive environment conducive to proper engagement [MKKS18].

Identifying a research gap, recent studies have highlighted the untapped potential of deep learning-based, ethically grounded chatbot guidance across diverse domains. Despite extensive research in influencing and technology separately, a specific gap exists in leveraging deep learning algorithms to enhance the usage of chatbots [ZOL+20]. The proposed approach aims to bridge this gap by exploring the application of deep learning techniques for ethical guidance. We aim to move with distinction from the existing literature on influencing by utilizing only certain neutral traits (like statistical support, inspiration appeals, rational appeals, and local/government initiatives) and removing the negative influencing techniques (like power dynamics, coercion, etc.) in our guidance system. By leveraging these methods, the intention is to empower chatbots with enhanced capabilities, opening new avenues for providing efficient information in various domains.

Chapter 9

Dataset

9.1 Dataset Creation

In creating our EcoNudge dialogue dataset aimed at promoting environmental sustainability through a conversational agent, a pivotal emphasis was placed on persona-based categorization. Initially, we facilitated human-human interactions by presenting a curated list of environmental suggestions to diverse participants. This initial phase allowed us to observe and categorize users based on their responses, identifying distinct user personas. Analyzing these interactions while referencing existing literature [Cia03], we classified users into five broad personas categories: Active, Skeptical, Inquisitive, Materialistic, and Biased. Each persona encapsulates unique traits, beliefs, and cognitive biases regarding environmentally conscious behavior.

Understanding user personas plays a crucial role in shaping subsequent dialogue flows. Each persona presents distinct challenges and requires tailored guidance strategies. For instance, the active persona demonstrates an inherent inclination towards environmentally conscious actions, whereas the skeptical persona exhibits a predisposition towards skepticism and resistance. The Inquisitive persona seeks detailed information, the materialistic persona is motivated by tangible benefits, and the biased persona exhibits strong predispositions or resistance against environmental suggestions.

Crafting convincing arguments for each suggestion includes diverse supporting statements spanning benefits, stats, anecdotes, and FAQ-style information, specifically curated to resonate with each persona's inclinations and biases. Additionally, generating natural and diverse dialogue flows between users and agents factored in the personas' cognitive biases and receptivity levels, ensuring authenticity and effectiveness in conversational interactions.

9.1.1 User Personas

The data users were categorized into five personas based on the initial human-human interactions and the literature survey [Cia03, Cia16, WSK+19]. Table 9.1 contains a description of various user personas.

Persona	Description
Active	Embracing a proactive approach, the active persona ardently engages with environmentally conscious behaviors, consistently seeking novel methods to contribute positively to the environment.
Skeptical	The skeptical persona demonstrates a persistent sense of doubt towards environmental initiatives. They are cautious, often requiring substantial evidence or reassurance to consider altering their existing behaviors or habits.
Inquisitive	Characterized by a deep curiosity, the inquisitive persona actively seeks comprehensive information and an in-depth understanding of environmental practices. Their inclination towards asking probing questions highlights their eagerness to know more.
Materialistic	Driven by tangible benefits and practical gains, the materialistic persona emphasizes eco-friendly choices' economic or personal advantages. They prioritize practical, tangible benefits.
Biased	The biased persona exhibits firm preconceived notions or entrenched resistance towards environmental recommendations. Overcoming these biases might require strategies tailored to address their concerns or misconceptions.

Table 9.1 User Personas

9.1.2 Bot/Agent Actions

Based on the user persona and input, the agent decides and responds using the dialogue-acts outlined in Table 9.2. We derive dialogue through observation (during the human-human interaction phase) and literary analysis [Cia03, Cia16, WSK+19]. Revision, feedback, collaboration, genre conventions, and language/style also shape dialogue. These methods ensure dialogue authenticity and effectiveness in serving narrative, character, and thematic purposes.

Action	Description
Benefits	Agent explains the benefits associated with the suggestion.
InspirationalAppeal	Agent employs an inspirational appeal to support the suggestion.
Consensus	Agent provides information or examples to establish a consensus.
RationalAppeal	Agent appeals to reason and logic to support the suggestion.
Action	Agent suggests possible actions to support sustainability.
inquiry	Agent responds to a user's inquiry or question.
SocialAppeal	Agent makes a social appeal to support the suggestion.
Encouragement	Agent offers encouragement or motivation to the user.
Intrigue	Agent acknowledges the user's enthusiasm regarding positive results.
Support	Agent provides support through data, references, or stories.
Openness	Agent responds positively to the user's enthusiasm for more suggestions.
Follow-up	Agent asks for updates on the previous suggestion or action.
Suggestion	Agent provides a new suggestion or idea.
Addressdoubt	Agent addresses doubt or uncertainty expressed by the user.
Addresscomplain	Agent addresses a specific complaint or concern raised by the user.
HopefulConsidersation	Agent expresses hope that the user will succeed when trying the suggestion.
Simplify	Agent simplifies a complex or hard-to-follow suggestion for better understanding.

Table 9.2 Bot actions and descriptions.

9.1.3 Dialogue Flow Generation

The potential conversation paths were derived from real human interactions, mimicking a user-agent dialogue. Each participant was assigned a specific persona, defining their level of curiosity toward the suggestion. This persona delineates the extent of information the user seeks before considering the suggestion, reflecting concerns about the environment or a focus on financial gains. Additionally, the persona guides the selection of suggestions offered to the user, aligning with their preferences and aversions. These simulated dialogues were annotated with a sequence of dialogue acts, encapsulating various conversational behaviors.

Furthermore, the usage of personas in this process was significant. It shaped the content

and direction of the conversations and guided the allocation of suggestions according to the user's inclinations and interests. The personas helped simulate diverse user behaviors, indicating varying degrees of openness, skepticism, environmental consciousness, or focus on tangible benefits. These personas served as a key factor in customizing the dialogue acts and, subsequently, the suggestion offerings, ensuring a tailored approach to meet simulated users' distinct needs and tendencies. Impacted by personas, this annotated sequence of dialogue acts formed the basis for training the prompt-based Language Model (LLM) for data.

9.1.4 Data Preparation and Validation

For conversation generation, we employ the Llama2 model, known for its consistent delivery of realistic and coherent dialogues among available language models. Using dialogue acts as cues, we guide the Llama2 model's conversation creation process, providing essential directives to structure and steer the dialogues. This integration of prompts facilitates purposeful and organized conversations closely aligned with our intended conversational objectives. To ensure coherence and relevance, we employ a prompting technique. Our prompts, including a [SYSTEM] component outlining the conversation's context and objectives, coupled with dialogue acts, direct each conversational turn. Additionally, the Llama2 model receives supplemental suggestion-related information alongside persona details and chosen suggestions. This diverse input approach results in contextually pertinent and compelling conversations, enhancing overall coherence and relevance within the dataset.

Subsequently, our validation procedure involves linguistics and computational linguistics experts scrutinizing dialogues for fluency, coherence, and alignment with dialogue flows and personas.

9.1.5 Dataset Statistics

The dataset consists of 400 evenly distributed interactions, with 80 interactions dedicated to each predefined user persona. Each interaction comprises five conversations, each depicting

multiple instances of communication between the user and the agent. The design of these interactions reflects the understanding that guidance behavior typically occurs gradually over time. The system structure accommodates this gradual process by organizing the dialogues accordingly. Specifically, the gap between two consecutive conversations within the same interaction signifies the time allocated for the user to experiment with the suggestions or serves as intervals before introducing new suggestions. This mimics a scenario where the user requires time or pauses between interactions to effectively implement or consider the given tips. Consequently, the dataset comprises a total of 2000 dialogues, with an average dialogue length of ten turns.

Dataset Information	
Total Communications	400
Communications per User Persona	80
Total Conversations	2000
Average Dialogue Length	10 turns

Table 9.3 Dataset Statistics

Chapter 10

Models and Methodology

10.1 Models

In our experiments, we utilize several advanced language models to assess their effectiveness:

Llama2 (Large Language Model Meta AI): Developed by Meta AI, Llama2 is a foundational Large Language Model (LLM) pre-trained on an extensive dataset comprising 2 trillion tokens. Due to its popularity and wide usage in recent work, we select Llama2 to represent the available LLMs in our study.

GPT-2: GPT-2, developed by OpenAI, is a Transformer-based model known for its ability to generate coherent text. It has undergone pre-training on a vast corpus of English text using self-supervised learning. GPT-2 utilizes a masking mechanism to predict tokens based on preceding sequences during training. We choose GPT-2 as a baseline model for comparison and as the base model for our reinforcement learning-based fine-tuning.

GODEL: GODEL is a large-scale pre-trained model specifically tailored for goal-directed dialogues. It is parameterized with a Transformer-based encoder-decoder architecture, enabling it to generate responses grounded in external textual knowledge. GODEL leverages additional text information to produce contextually relevant responses in dialogues. As a baseline representative of the sequence-to-sequence transformer models, GODEL offers a different approach compared to GPT-2's autoregressive decoder-based approach.

10.2 Methodology

In our system, we ensure that the language model follows our suggested strategies when formulating responses. Initially, we fine-tune a pre-trained language model, GPT-2-medium, in a supervised setting using traditional cross-entropy loss between the ground truth and predicted utterances' probability distributions.

In our current work, we utilize six reward functions to evaluate the performance of the generated responses. These reward functions are categorized into two types: generic rewards, which deal with the clarity of response concerning the dialogue setting, and task-specific rewards, which correspond to incorporating the strategies above into the model.

The first reward function, called the Length Reward, penalizes the generated response by considering the absolute difference between its length and the target length. This mechanism incentivizes the model to produce responses that closely align with the desired length, fostering the generation of concise and coherent replies. The formula is expressed as follows:

$$\mathrm{R(Len)} = -\left|L - T\right| \tag{10.1}$$

Here, L denotes the length of the generated response, and T represents the target length.

To ensure the quality of generated responses, preventing instances where highly rewarded replies lack grammatical correctness or coherence, we employ the Semantic Coherence Reward method. This involves assessing the mutual information between the action a and preceding dialogue turns in the history to ascertain the coherence and appropriateness of generated responses. The equation for the reward function $R(SC)$ is formulated as follows:

$$R(SC) = \frac{1}{N_a} \log p_{\mathrm{s2s}}(a|q_i, p_i) + \frac{1}{N_{q_i}} \log p_{\mathrm{s2s}}(q_i|a) \tag{10.2}$$

Here, $p_{\mathrm{seq2seq}}(a|p_i, q_i)$ represents the probability of generating response a given the previous dialogue utterances $[p_i, q_i]$. Meanwhile, $p_{\mathrm{backward\ seq2seq}}(q_i|a)$ denotes the backward probability of generating the prior dialogue utterance q_i based on response a. The training of

$p_{\text{backward seq2seq}}$ follows a similar procedure as standard models but with sources and targets interchanged. To mitigate the impact of target length, both $\log p_{\text{seq2seq}}(a|q_i, p_i)$ and $\log p_{\text{backward seq2seq}}(q_i|a)$ are scaled by the length of targets.

We also use four reward functions to assess and guide the user properly:

1. Change in Opinion Sentiment Reward: This function evaluates the change in sentiment from an initial state to a final state. It assigns a zero reward when the sentiment transition corresponds to specific predefined patterns, encouraging the model to recognize and appropriately respond to significant changes in opinions or sentiments.

2. Intermediate Completion Reward: This function evaluates the reward based on different actions taken, such as rejecting, considering, or implementing suggestions. The reward value depends on the user's acceptance of the suggestion.

3. Positive Affirmation Reward: This function assesses users' responses to their liking or disliking of the suggestion. It assigns a reward of zero for a user response indicating dislike and a reward of one for a response affirming usefulness.

4. Personalized Support Reward: This function combines multiple reward components based on different support types. The overall reward is the sum of weighted rewards for each support type and a diversity score that measures how evenly the agent utilizes different support types.

Chapter 11

Evaluation Setup

In our automatic evaluation, we employ several metrics to assess different aspects of the generated responses. For syntactic evaluation, we use the following metrics:

- **BLEU** (Bilingual Evaluation Understudy): Measures the n-gram overlap between the generated response and the reference response.
- **METEOR** (Metric for Evaluation of Translation with Explicit ORdering): Computes the harmonic mean of precision and recall using explicit word-to-word alignment.
- **ROUGE-L** (Recall-Oriented Understudy for Gisting Evaluation-Longest-common-subsequence): Measures the longest common subsequence between the generated and reference responses.
- **CIDEr** (Consensus-based image description evaluation): Calculates the consensus among multiple reference responses.
- **SPICE** (Semantic propositional image caption evaluation): Evaluates the semantic similarity between the generated and reference responses.

For semantic similarity evaluation, we utilize the following metrics:

- **Skip-Thought Cosine Similarity (STCS)**: Measures the cosine similarity between the skip-thought embeddings of the generated and reference responses.

- **Greedy Matching Score (GMS)**: Computes the optimal matching between the words in the generated and reference responses.
- **Vector Extrema Cosine Similarity (VECS)**: Measures the cosine similarity between the vector extrema embeddings of the generated and reference responses.
- **Embedding Average Cosine Similarity (EACS)**: Computes the average cosine similarity between the word embeddings of the generated and reference responses.

In our human evaluations, three proficient postgraduate evaluators interacted with the proposed system 100 times to assess its performance. They rated the system on a scale of 1 to 5 for three task-specific criteria:

- **Support Consistency (S-Con)**: Measures the consistency of the negotiation approach within a dialogue.
- **Guidance Efficacy (G-Eff)**: Evaluates the ability of the system to present compelling arguments, reasoning, or incentives that guide the other party.
- **Dialogue-Engagingness (D-E)**: Assesses the extent to which a conversation or dialogue is interesting, captivating, and can hold the participants' attention.

Additionally, the annotators evaluated the system on two general human evaluation metrics:

- **Fluency**: Measures the smoothness and naturalness of the generated responses.
- **Context Relevance**: Assesses the relevance of the generated responses to the context of the conversation.

Chapter 12

Results and Analysis

We evaluate five models: GPT2-rl, GPT2, GODEL, Llama2-prompted, and Llama2-finetuned. GPT2-rl is our proposed RL-based model, while GPT2 and GODEL serve as fine-tuned baselines. Llama2-prompted directly prompts the available Llama variant, whereas Llama2-finetuned is its fine-tuned version.

Models	Bleu_1	METEOR	ROUGE_L	CIDEr	SPICE	STCS	EACS	VECS	GMS
GPT2	0.193	0.116	0.14	0.133	0.096	0.598	0.882	0.462	0.73
GODEL	0.275	0.159	0.238	0.487	0.2	0.721	0.916	0.509	0.76
Llama2-prompted	0.23	0.145	0.168	0.175	0.162	0.657	0.854	0.483	0.745
Llama2-finetuned	0.379	0.217	0.323	0.875	0.306	0.826	0.942	0.585	0.807
GPT2-rl	0.349	0.202	0.313	0.875	0.276	0.811	0.937	0.569	0.795

Table 12.1 Results for generic automatic evaluation of the proposed **EcoNudge** dialogue dataset

In automatic evaluation (Table 12.1), GPT2 performs weaker in syntax but compensates with decent semantic similarity. GODEL consistently outperforms GPT2 across various metrics, indicating better syntactic quality and semantic balance. Llama2-prompted falls between GPT2 and GODEL due to static prompt usage. Llama2-finetuned achieves highest scores, with GPT2-rl closely following despite significant parameter difference. Human evaluation involved 100 interactions, where evaluators rated systems on Support Consistency, Guidance Efficacy, and Dialogue-Engagingness. Llama2-finetuned excelled in consistency

and engagement, while GPT2-rl and Llama2-finetuned led in guidance. General metrics like Fluency and Context Relevance were also considered, with Llama2-finetuned scoring highest in context relevance. The proposed model exemplifies effective guidance strategies, including adjusting recommendations based on user feedback and persona updates. These strategies were validated during human-human conversations and reflected in the annotated dataset. The inter-annotator agreement among human evaluators was 80.5%.

Models	Fluency	CR	S-Con	G-Eff	D-E
GPT2	3.7	3.19	2.41	2.7	3.3
GODEL	4.1	3.58	3.24	3.1	3.7
Llama2-prompted	4.4	2.68	1.61	2.6	3.4
Llama2-finetuned	4.4	4.4	3.75	3.6	4.3
GPT2-rl	4.6	4.66	3.77	3.9	3.7

Table 12.2 Results for generic Human evaluation of the proposed **EcoNudges** dialogue dataset

Chapter 13

Conclusion

Our study introduces a novel approach to utilize chatbots as guides for promoting environmental sustainability. We have created the EcoNudge dialogue dataset, a valuable resource that enables the analysis of various user personas' responses to sustainability-focused guidance techniques. Our main focus was on developing a reinforcement learning model that is aware of user personas, allowing chatbot agents to tailor guidance strategies according to individual traits.

Through a combination of theoretical insights and practical implementations, our research bridges the gap between the urgent need for environmental conservation and effective ways to engage society. Our findings indicate that finely-tuned smaller models can outperform well-prompted Large Language Models (LLMs). Additionally, employing task-optimized reinforcement learning further enhances performance, bringing it closer to that of finely-tuned LLMs.

References

[Cia03] Robert B Cialdini. *Influence*. Influence At Work, 2003.

[Cia16] Robert Cialdini. *Pre-suasion: A revolutionary way to influence and persuade*. Simon and Schuster, 2016.

[CWW+23] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*, 2023.

[DG20] Irina Dokukina and Julia Gumanova. The rise of chatbots–new personal assistants in foreign language learning. *Procedia Computer Science*, 169:542–546, 2020.

[GSS+13] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. Policy shaping: Integrating human feedback with reinforcement learning. *Advances in neural information processing systems*, 26, 2013.

[LYJ+23] Xiang Li, Yiqun Yao, Xin Jiang, Xuezhi Fang, Xuying Meng, Siqi Fan, Peng Han, Jing Li, Li Du, Bowen Qin, et al. Flm-101b: An open llm and how to train it with $100k budget. arXiv preprint arXiv$: 2309.03852, 2023.

[MKKS18] Robert R Morris, Kareem Kouddous, Rohan Kshirsagar, and Stephen M Schueller. Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. *Journal of medical Internet research*, 20(6):e10148, 2018.

[NHX+23] Erik Nijkamp, Hiroaki Hayashi, Caiming Xiong, Silvio Savarese, and Yingbo Zhou. Codegen2: Lessons for training llms on programming and natural languages. *arXiv preprint arXiv:2305.02309*, 2023.

[Par63] Talcott Parsons. On the concept of influence. *Public opinion quarterly*, 27(1):37–62, 1963.

[Rav08] Bertram H Raven. The bases of power and the power/interaction model of interpersonal influence. *Analyses of social issues and public policy*, 8(1):1–22, 2008.

[Sch17] Christian Schubert. Green nudges: Do they work? are they ethical? *Ecological economics*, 132:329–342, 2017.

[SH90] Chester A Schriesheim and Timothy R Hinkin. Influence tactics used by subordinates: A theoretical and empirical analysis and refinement of the kipnis, schmidt, and wilkinson subscales. *Journal of applied psychology*, 75(3):246, 1990.

[WSK+19] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *arXiv preprint arXiv:1906.06725*, 2019.

[Zam17] Jennifer Zamora. Rise of the chatbots: Finding a place for artificial intelligence in india and us. In *Proceedings of the 22nd international conference on intelligent user interfaces companion*, pages 109–112, 2017.

[ZOL+20] Jingwen Zhang, Yoo Jung Oh, Patrick Lange, Zhou Yu, and Yoshimi Fukuoka. Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet. *Journal of medical Internet research*, 22(9):e22845, 2020.

[ZZL+23] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.