Analysing Patterns of Repetition in Child-Adult Dialogues

Maryam Bala

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Master of Science in Artificial Intelligence
of the
University of Aberdeen.

Department of Computing Science

2024

Declaration

No portion of the work contained in this document has been submitted in support of an application for a degree or qualification of this or any other university or other institution of learning. All verbatim extracts have been distinguished by quotation marks, and all sources of information have been specifically acknowledged.

Signed: *maryambala*

Date: 03 May 2024

Abstract

From the earliest babbles to the eloquent discourse of adulthood, First Language Acquisition (1LA) is a complex journey that has captivated researchers for decades. At the heart of this journey lies the dynamic relationship between children and their caregivers. Caretakers often use repetition in their interactions with children, serving to reinforce their message or capture attention . This input from caregivers has been shown to significantly aid children in acquiring their native language. This study investigates the repetitive strategies used as feedback mechanisms within child-adult dialogues, and explores the potential of Large Language Models (LLMs) for both detecting and generating these strategies.

By utilising data from the Child Language Data Exchange System (CHILDES), a rule-based and an automatic detection method, which utilises perplexity and cosine similarity properties derived from LLMs, are employed for feedback detection. The automatic detection classifier achieved high accuracy, indicating that LLM properties can serve as valuable features for feedback detection. Three state-f-the-art LLMs - OPT, DialoGPT and Gemma are fine-tuned on a large corpus of child-adult dialogues to generate feedback responses. The three models perform satisfactorily when assessed using both automatic metrics and human evaluation, with OPT performing slightly better than DialoGPT and Gemma across both evaluation metrics. The findings from this research introduce a novel approach of using properties of LLMs to detect feedback categories in child-adult dialogues. Furthermore, the generated responses from the fine-tuned LLMs demonstrate their applicability in dialogue generation systems for providing tailored feedback responses to children in 1LA.

Acknowledgements

I am indebted to the Petroleum Technology Development Fund (PTDF) for their sponsorship, without which this Master's Degree would not have been possible. I am sincerely grateful for this opportunity.

I extend my deepest gratitude to my supervisor, Dr. Arabella Jane Sinclair, for her unwavering support and invaluable guidance throughout my research. Her expertise in Natural Language Processing has been instrumental in shaping the understanding necessary for completing this thesis. Dr. Sinclair's patience, willingness to assist, and encouragement have been immeasurable, and I am profoundly thankful to her.

I would also like to express my sincere appreciation to Dr Johannes Heim. This project would not have been possible without his invaluable expertise and insights in the field of Linguistics. His guidance and input have played a crucial role in shaping the outcome of this project.

Finally, I would like to express heartfelt appreciation to my family and friends for their unwavering support throughout this program.

Contents

List of Figures

List of Tables

List of Algorithms

Listings

Chapter 1

Introduction

First language acquisition (1LA) is the process by which infants acquire their native language (Clark and Casillas, 2015). It is a fundamental milestone in human development, marking the emergence of linguistic competence from infancy to early childhood. Repetition in dialogues, play a central role in this process, as caregivers frequently use it as a form of feedback to reinforce language learning (Lester et al., 2022). Through repeated exposure, children internalize linguistic patterns and structures (Clark, 2020). The study of child-adult dialogues offers a unique perspective on the dynamic processes of 1LA, providing valuable insights into how children learn to communicate and interact with their caregivers. To better understand the complex nature of these interactions, researchers have relied on the Child Language Data Exchange System (CHILDES) (MacWhinney, 2000). This extensive database, containing transcripts of conversations between children and their caregivers across various languages and contexts, has become a valuable resource for studying language development.

Large Language Models (LLMs) have become mainstream since the release of ChatGPT (Radford et al., 2019), due to their ability to generate human like text in a vast array of topics (Brown et al., 2020; Bommasani et al., 2021). At their core, these models leverage deep learning techniques to analyze massive amounts of text data, identifying patterns that allow them to understand and generate natural language (Brown et al., 2020). Feedback in language learning, the process of providing constructive guidance in response to linguistic input, is significant in enhancing language learning experiences (Ferreira and Atkinson, 2008). Taking advantage of LLMs' ability to analyze and generate text, innovative approaches can be explored for feedback generation tailored to individual learners' needs. This approach offers a nuanced understanding of how feedback influences language acquisition processes and can inform the design of more effective language learning interventions.

The primary aim of this project is to detect various categories of caregiver feedback present in child-adult dialogues, and to harness the capabilities of LLMs for both detecting and generating such feedback. The intent is to develop methods for accurately detecting different feedback strategies through a large scale analysis of interactions between children and caregivers from the CHILDES database. Following successful detection, the study then explores the potential of LLMs in generating contextually relevant feedback responses. The ability to detect and generate appropriate feedback is crucial for creating adaptive and engaging conversational agents that can support and enhance the language acquisition process for young learners. Ultimately, the goal of this research is to advance the understanding of feedback mechanisms present in child-adult dialogues and make use of cutting-edge language models to explore solutions that can positively impact the field of 1LA and education.

1.1 Motivation

This project seeks to develop tools that aid linguistics researchers in corpus analysis, while also devising methods to contribute to the development of future conversational systems tailored for children. Gaining insights into child language and feedback patterns can deepen the understanding of language acquisition in children, which can subsequently inform child developmental pedagogy. Simultaneously, this knowledge can also benefit the design and creation of more user-friendly, adaptive, and age-appropriate dialogue agents for children, as observed in the case of robotic toys (Marti and Giusti, 2010; O'Brien et al., 2021). This is an area that can potentially benefit from incorporating novel large language model technology. However, to fully optimize the potential of language models in this context, interdisciplinary research such as this is essential to fully understand how this mechanism works. To maintain a focused scope, this project concentrates on adult feedback, specifically question and answer sequences, between children and their caregivers. By utilizing the power of advanced language models and leveraging interdisciplinary insights, the aim is to contribute to the field of first language acquisition.

1.2 Research Objectives

The research aims to answer the following research questions:

- **Research Question 1:** ***Feedback Detection*** Can diverse feedback strategies employed in child-adult interactions be effectively detected ?

- **Research Question 2:** ***Feedback Generation*** Can contextually appropriate feedback be generated for children in natural language interactions by utilizing LLMs ?

Consequently, the research will focus on the following objectives:

1.2.1 Feedback Detection

For feedback detection, two methods are employed: a baseline approach that follows Hiller and Fernández (2016) by primarily using vocabulary overlap as a primary heuristic for identifying repetitions in child-adult interactions and a novel approach that leverages properties of LLMs specifically perplexity and cosine similarity, to investigate their potential as features for aiding feedback detection.

1.2.2 Feedback Generation

Once instances of feedback have been identified through the detection approach, the next objective focuses on investigating the capability of LLMs in generating appropriately human-like feedback utterances in a given dialogue context. This objective assesses the model's proficiency in not only detecting errors but also contributing meaningfully to the ongoing conversation.

1.3 Contributions

This project makes several notable contributions to the field of 1LA and Natural Language Processing (NLP), fostering interdisciplinary collaboration and knowledge exchange. The integration of these fields opens up new avenues for research and innovation in supporting children's language development. The main contributions are as follows:

- The project introduces a novel approach to identifying feedback in child-adult dialogues by combining rule-based algorithms with the properties of LLMs and demonstrates reasonable performance, even with a very small dataset, the highest of which makes use of the heuristics derived from LLMs.

- The project explored the potential of a suite state-of-the-art language generation models for generating caregiver feedback. By adapting these models to the context of child-adult dialogues, this research demonstrated their effectiveness in capturing the nuances and patterns of feedback in naturalistic interactions. The insights gained from this portion of the research holds the potential to guide the strategic design and use of LLMs within conversational educational environments for children.

1.4 Thesis Structure

The dissertation is divided into 6 chapters. Namely:

- Chapter 1: **Introduction** provides an overview of the research topic, the motivation behind the study, and the main objectives and contributions of the project.

- Chapter 2: **Background and Related Work** reviews the existing literature and approaches relevant to the research topic, identifying gaps and opportunities for further investigation.

- Chapter 3: **Methodology** describes the research design, data collection and pre-processing techniques, the proposed approach, and the evaluation metrics used in the study.
- Chapter 4: **Results** presents the experimental results.
- Chapter 5: **Analysis and Discussion** analyzes the findings, and discusses their implications in relation to the research questions and objectives.
- Chapter 6: **Conclusion and Future Work** summarizes the main contributions and insights gained from the study, addresses the limitations, and outlines potential directions for future research in the field

Chapter 2

Background and Related Work

This chapter provides a brief overview of the foundational concepts and prior research that underpin the current study. The discourse will cover essential elements of first language acquisition, feedback strategies in child-caregiver interactions and text generation using LLMs. An overview of current research on identification of feedback in child adult dialogues and applications of feedback in text generation systems will also be provided.

2.1 Background

This study aims to enhance the understanding of first language acquisition through the investigation of different types of input from their caretakers. Additionally, the study intends to make a contribution to the domain of natural language processing by investigating the use of language models for feedback generation, which can be beneficial for conversational agents. To achieve these goals, fundamental concepts such as first language acquisition, repetition in child language learning, and large language models application in text generation will be examined in this section.

The section is divided into three main segments. The initial subsection examines first language acquisition, covering theories of language acquisition and the role of linguistic input. The second sections provides an overview of repetition as a form feedback in 1LA. The third subsection provides an overview of large language models and their application in generating contextually appropriate text for dialogues.

2.1.1 First Language Acquisition

First Language Acquisition (1LA) is a complex journey that begins right from birth and continuous through the early years of life. It refers to the process by which humans learn and understand language in an instinctive way without any instruction at the early years of life (Clark and Casillas, 2015).

2.1.1.1 Theories of Language Acquisition

Numerous theories have been proposed on how children first learn language. However, this brief discussion will focus on two prominent theories, each presenting a distinct perspective:

- **Nativist Theory:** Chomsky (1965) proposed the existence of an innate language acquisition device (LAD) in humans, facilitating the process of language acquisition. According to this, children possess inherent familiarity with linguistic structures and principles which direct their language acquisition. Nativism focuses on the formal aspects of language structure, syntax, and grammar as central to language acquisition (Chomsky, 1965). This theory proposes that children possess an an innate Universal Grammar (UG) that allows them to discern the structure of their native language through mere exposure to it.

- **Usage-based Theory:** Introduced by Tomasello (2005), this theory emphasizes that language structure emerges from language use and that children build their language skills based on their general cognitive abilities. While distinct from Behaviorism (Skinner, 1957) in many aspects, this theory shares some roots in Behaviorist principles which suggests that language is mostly learned through reinforcement and imitation. According to the usage-based theory, children learn language by being exposed to linguistic input from caregivers and other language users (Tomasello, 2005). The frequency and quality of this input play a significant role in shaping children's language skills and abilities, recognising the importance of feedback in 1LA. Feedback from caregivers and more proficient language users provides children with valuable reinforcement, corrections, and models for refining their language skills (Tomasello, 2005).

Despite the differences between the nativist and usage-based theories, both recognize the interplay between innate linguistic abilities and environmental language input in shaping language development. They both acknowledge that language acquisition is a complex process that involves a combination of innate abilities and exposure to linguistic data.

2.1.1.2 Linguistic Input and Language Acquisition

The significance of linguistic input and its role in language acquisition has been a focal point of discussion, characterized by debates surrounding the poverty of the stimulus and infant-directed speech (IDS), both of which have provided significant perspectives.

Poverty of the stimulus (Chomsky, 1965) highlights the gap between the linguistic input children receive and the complex language system they eventually acquire. This suggests that the input received by children alone is inadequate to explain how rapid and efficient learning occurs, considering the complex and intricate nature of linguistics structures (Laurence and Margolis, 2001). This phenomenon has resulted in various theoretical viewpoints regarding the acquisition of language.

However, IDS, a distinct form of speech directed towards infants, has shown that the input children receive is not lacking in quality but instead rich in linguistics and prosodic features which may facilitate the acquisition of language (Fernald et al., 1989; Saint-Georges et al., 2013). IDS, also referred to as parentese or motherese, is a mode of communication distinguished by elevated pitch, reduced tempo, amplified intonation, and streamlined grammar and vocabulary utilized by adults when interacting with young children (Ferjan Ramírez et al., 2020). In the study by Ferjan Ramírez et al. (2020), it was observed that implementing parent coaching aimed at enhancing the use of parentese and increasing conversational interactions resulted in notable improvements in language skills among infants. The infants who received coaching demonstrated higher rates of vocalizations, more rapid development in conversational exchanges, and a more extensive range of words in comparison to the control group. The linguistic stimuli that children are exposed to, particularly when supplemented with parentese and interactive conversations, can greatly enhance the process of language learning (Ferjan Ramírez et al., 2020).

2.1.2 Repetition in Language Acquisition

Repetition is a fundamental element in child-adult dialogues. Caregivers instinctively engage in this process, providing children with opportunities to hear and practice language structures. Several studies have explored the different ways caregivers utilize repetition as a form of feedback and its effectiveness in fostering children's language development. One influential work in this area is by Demetras et al. (1986) who defined two key types of feedback - explicit and implicit.

- Explicit feedback involves direct comments on the correctness of a child's utterance. This can include expressions of approval, disapproval, or corrections.

- Implicit feedback, on the other hand, focuses on encouraging communication and fostering understanding. It can take the form of requests for clarification, repetitions, or conversational responses.

Building on this, Chouinard and Clark (2003) further investigated the role of reformulations, a type of implicit feedback where adults rephrase children's erroneous utterances. Through detailed analyses of child-parent conversations, they demonstrated how reformulations help children recognize the contrast between their own utterances and the target forms. Work by Saxton (2000) and Saxton et al. (2005) focused on the concept of negative evidence, which refers to feedback indicating that a child's utterance is incorrect. They argue that negative evidence, whether explicit or implicit, is essential for children to eliminate grammatical errors and refine their language skills.

Recent studies have continued to shed light on the nuances of feedback in 1LA. Hiller and Fernández (2016) used the term corrective feedback, defined as a type of feedback that is provided in response to a child's errors on utterances. Their findings indicated that corrective feedback plays a significant role in predicting learning outcomes. This influence becomes apparent after a delay of around 9 months, implying that feedback has a lasting effect on learning over time, extending beyond immediate responses from the child. Nikolaus and Fourtassi (2023) studied a form of feedback referred, to as communicative feedback, a process that occurs when children and their caregivers strive to establish a shared understanding. The researchers argue that feedback is not only a way for children to learn about the linguistic conventions of their community but also a means for them to learn about the communicative intentions of their interlocutors. This perspective highlights the importance of considering the social and pragmatic aspects of feedback in child language learning.

This section has highlighted the diverse terminology researchers have used to describe the input caregivers give to children to shape language development. These terms often reflect different aspects of the interaction and the specific focus of the research. While these definitions have subtle differences, they all refer to the various strategies caregivers employ to support 1LA. In this research, the term feedback will be used broadly to encompass these diverse range of caregiver responses, including both implicit and explicit forms. However, the primary focus will be on implicit negative feedback, the more subtle strategies caregivers use that do not involve explicit correction.

2.1.3 Text Generation with Large Language Models

Large Language Model (LLMs) have significantly changed the landscape of NLP in recent years. Commonly based on the transformer architecture, introduced by Vaswani et al. (2017), they are models with hundreds of billions (or more) of parameters, which undergo training on extensive text datasets (Zhao et al., 2023). The transformer architecture uses self-attention mechanisms to process input sequences and capture long-range dependencies in datasets (Vaswani et al., 2017). This allows LLMs to effectively handle context and generate coherent and contextually relevant outputs. LLMs are pre-trained by employing self-supervised learning methods, such as masked language modeling (MLM) and next sentence prediction (Devlin et al., 2018). These specified objectives empower the models to acquire comprehensive language representations that can be fine tuned for specific downstream tasks with limited labeled data. Notable examples of LLMs include GPT models (Radford et al., 2018; Brown et al., 2020), BERT (Devlin et al., 2018), Open Pretrained Transformer (Zhang et al., 2022), Galactica (Taylor et al., 2022), PaLM (Chowdhery et al., 2023), LLaMA (Touvron et al., 2023) and Gemma (Gemma et al., 2024).

Some key applications of LLMs in NLP include, language generation, text classification, question answering, machine translation, sentiment analysis and many more (Brown et al., 2020; Bommasani et al., 2021). Language generation stands out as a crucial application of LLMs, allowing for the production of text resembling human language. This capability has opened up possibilities for applications like story generation, conversational interfaces, and content creation (Radford et al., 2018; Brown et al., 2020). Recent advancements have opened up opportunities for the creation of conversational systems capable of producing customized, contextually relevant responses designed for users. However, it is still a challenging task, as it requires capturing the unique linguistic styles, preferences, and knowledge of individual speakers (Li et al., 2016). Personalized dialog systems are designed to offer a richer, more organic, and efficient communication experience by adjusting to the distinct traits, choices, and requirements of every individual (Zhang et al., 2018; Qian et al., 2017).

This research leverages the language generation capabilities of LLMs to produce contextually appropriate responses in dialogues between children and adults. By fine-tuning LLMs on a dataset of child-adult interactions containing instances of feedback, the models are trained to generate feedback tailored to the child's linguistic competence and the conversational context.

2.1.3.1 Evaluation of Text Generation Systems

The evaluation of text generation systems holds significant importance in determining the quality and effectiveness of generated texts. Automatic metrics such as BLEU (Papineni et al., 2002), Bert (Zhang et al., 2019a), Mauve (Pillutla et al., 2021) and perplexity (Serban et al., 2016), enables a quantitative evaluation of generated responses. However, these metrics might not fully capture the subtleties of personalisation and user-specific appropriateness. Human assessment, conducted through user studies or expert annotations, is frequently essential for evaluating the subjective elements of user-specific dialogue quality, such as engagement, relevance, and user satisfaction Liu et al. (2016).

Generating and evaluating appropriate texts is particularly crucial for dialogue agents intended to interact with children. The language used when engaging with children differs from that used for adults. Age and linguistic proficiency level must be taken into account when producing dialogue tailored for children. This study utilizes a combination of automated metrics and human assessment to evaluate the feedback generated using LLMs.

2.2 Related Work

This section comprises of two sub sections. In the first sub section, an analysis will be conducted on the current research pertaining to the detection of feedback within interactions between children and adults. The second sub section will focus on investigating the existing literature on the generation of feedback in dialogues as well as the evaluation metrics and methodologies used to assess the quality and effectiveness of generated texts.

2.2.1 Feedback Detection in Child-Adult Dialogues

The role of caregiver feedback in facilitating 1LA has been extensively studied. Researchers have long recognized the significance of the linguistic input and interactive exchanges between children and their caregivers as crucial factors that shape language learning (Hoff, 2006; Rowe, 2012). Early research into caregiver feedback primarily focused on analyzing feedback through observational studies and small-scale data (Snow, 1972; Saxton, 2000; Chouinard and Clark, 2003). These studies provided insights into different types of feedback, such as explicit and implicit corrections (Demetras et al., 1986), reformulations (Chouinard and Clark, 2003), and corrective feedback (Hiller and Fernández, 2016).

The automatic detection of feedback in child-adult conversations is an emerging area of research that combines techniques from linguistics and NLP. While research on this topic is scarce, Hiller and Fernández (2016) utilized a data-driven approach to investigate the impact of corrective feedback on 1LA. One significant contribution of the paper was analyzing a large dataset from the CHILDES Database (MacWhinney, 2000) using lexical overlap as a heuristic for extracting of corrective feedback pairs. The researchers then develop classifiers to detect subject omission errors and corrective feedback. The methodology used showed the potential of NLP approaches to facilitate large-scale analyses of child-adult data and provided valuable insights into the role of caregiver feedback in 1LA.

This thesis builds upon the work of Hiller and Fernández (2016) by first employing a similar rule based algorithm using lexical overlap to extract feedback from a large dataset of child-adult dialogues. However, it extends their methodology by incorporating features from LLMs, such as perplexity, to aid feedback identification. Whereas Hiller and Fernández (2016) focused on specific grammatical errors, this research takes a different approach, seeking to identify diverse forms of feedback through a large scale analysis of question-answer sequences between children and their caregivers. The combination of rule based and automatic feedback detection techniques using properties of LLMs as features and the examination of question-behavior related features offers a novel approach to identifying feedback in 1LA.

2.2.2 Feedback Generation

Some studies have demonstrated the potential of NLP to generate personalized feedback in Intelligent Tutoring systems (ITS) and found that generated feedback was effective in helping children improve their language skills and maintain engagement in the learning processe.e.g., (Ferreira and Atkinson, 2008; Kochmar et al., 2020; Troussas et al., 2023). In learning writing, Nagata (2019) pioneered the concept of generating automated feedback comments for writing, demonstrating its effectiveness for non-native learners with preposition errors. Building on this, Hanawa et al. (2021) explored various NLP techniques to generate feedback on a broader range of writing aspects. Rivers and Koedinger (2013) proposed a data-driven approach for automatic feedback generation in programming education, utilizing the program solution space to predict student progress and provide tailored feedback. Their approach aims to improve learning gains for novice programmers in large classes with limited instructor time. Loem et al. (2023) explored the effectiveness of using GPT-3 with prompt-based methods for Grammatical error Correction (GEC) tasks. The controllability of the model allowed for personalized language instruction, by providing tailored feedback to match individual proficiency levels. However, the study exclusively focused on GPT-3 and the reliance on quantitative metrics for evaluation, suggesting the need for broader investigations and additional qualitative analyses.

While there is limited research specifically focusing on generating feedback for children in first language acquisition, studies in related areas such as second language learning, programming education, GEC and writing feedback have demonstrated the potential of language models and NLG techniques for generating personalized and effective feedback. This study makes a contribution to this particular academic domain through the focus on providing feedback for children in child-adult conversations and conducting both an automatic and human evaluation of the generated responses.

Chapter 3

Methodology

3.1 Dataset

The primary source of data used for this research is the Child Language Data Exchange System (CHILDES) Database (MacWhinney, 2000) [1], a widely-used resource for child language acquisition research. The Database is a comprehensive collection of transcripts of child-adult dialogues in various languages, collected through a combination of naturalistic observations and structured experiments, with trained researchers carefully transcribing the verbal interactions between children and their caretakers. The extensive coverage of the database makes it an ideal resource for investigating child-adult interactions. This project will be utilizing corpora from the English Language section of the database. Table 3.1 shows a dialogue excerpt between a mother and child from the CHILDES database.

MOT:	look .
CHI:	a chocolate ball there .
MOT:	what can you see ?
CHI:	chocolate ball .
MOT:	a chocolate ball .
MOT:	think it's a picture of a football Thomas .
CHI:	football .

Table 3.1: Dialogue excerpt between a mother and child from the CHILDES database (MacWhinney, 2000)

3.1.1 Data Curation

Given the vastness of the CHILDES Database, a targeted approach is adopted to select a specific subset of data relevant to the research objectives. First, only dialogues involving children within the age range of 12 - 48 months (1 to 4 years) old are selected. From this collection, the focus

[1] This resource is available at https://childes.talkbank.org/access

was further narrowed to dialogues featuring a single child interacting with no more than 5 adults. An additional criterion is applied, if the number of days in age exceeded 15 for children aged 11 months, they are included in the age cutoff for 1 year. While there is an overlap with the data used by Hiller and Fernández (2016), certain dialogues are excluded due to the age and number of children filters applied to the data used for this research.

The resulting dataset comprises 2,026 dialogues with 46 different unimpaired, naturally developing children. In total, these dialogues have 2,050,530 utterances, with children contributing 776,163 of these utterances. The youngest age in the dataset is P0Y11M21D (11 months and 21 days) and the oldest age is P4Y00M15D (4 years and 15 days). Tables 3.2 and 3.3 provide an overview of the curated dataset.

Corpus	Start Age	End Age	Files per Child
Belfast	2;0	4;0	7
Bloom1970	1;8	3;1	8.3
Braunwald	1;5	3;11	173
Brown	1;6	4;0	49
Clark	2;2	3;2	45
Demetras1	2;0	3;11	26
Kuczaj	2;4	4;0	152
Lara	1;9	3;3	120
Manchester	1;8	2;11	33.9
Providence	1;0	4;0	58.2
Sachs	1;2	3;8	90
Snow	2;5	3;9	40
Suppes	1;11	3;3	51
Thomas	2;0	4;0	163
Weist	2;1	4;0	23.5

Table 3.2: Overview of Dataset

	Total	Average per Child
Dialogues	2,026	44.04
Utterances	2,176,057	44,576.74

Table 3.3: Number of dialogues and utterances

3.1.2 Data Pre-processing

The raw data is downloaded from the English Language section of the CHILDES website [2] in Extensible Markup Language (XML) format. To facilitate analysis, a custom Python script

[2] Available at https://childes.talkbank.org/data-xml/Eng-NA/ and https://childes.talkbank.org/data-xml/Eng-UK/

is developed to extract relevant information from the XML files based on the specified age range and participant criteria into a pandas data frame. Extracted information includes speaker role (child or adult), speaker name, speaker age, utterance text, and relevant metadata such as dialogue ID and corpora name.

3.2 Defining Feedback Categories

Previous research on caregiver feedback have primarily examined feedback responses, often in small-scale investigations (Demetras et al., 1986; Chouinard and Clark, 2003; Saxton, 2000; Saxton et al., 2005). What hasn't been fully explored yet is question-answer interactions between caregivers and children as a form of feedback. With this in mind, four distinct question answer sequence feedback types have been identified [3]:

3.2.1 Modelling Question Answer

This feedback type indicates an adult modeling how to answer a question. The adult initiates a question-and-answer sequence without any contribution from the child. In essence, this feedback type captures instances where an adult demonstrates question-answering behavior, potentially serving as a model for the child. Table 3.4 highlights an example of this sequence.

MOT:	where's Cromer ?
MOT:	oh Cromer's at home .

Table 3.4: Exampled of Modelling Question Answer

3.2.2 Exemplifying Question Answer

Exemplifying Question Answer refers to a situation where an adult reformulates or rephrases a child's question and then provides an answer. By reformulating the question and then providing an answer, the adult potentially aids the child's understanding of question-answering patterns. Table 3.5 shows an example of this.

CHI:	where Christmas cookies ?
MOT:	where are the Christmas cookies ?
MOT:	they're all gone .

Table 3.5: Example of Exemplifying Question Answer

[3] In this research, the identified feedback types are based on the work conducted by the Department of Linguistics the University of Aberdeen

3.2.3 Rephrasing Question

This feedback type involves an adult rephrasing or paraphrasing a child's question. It is considered a form of corrective feedback because it encourages the child's attempts at communication. The caregiver's rephrasing serves to provide a corrected version of the child's utterance and encourages the child to elaborate or explore their question further. A case of this feedback type is shown in Table 3.6.

CHI:	where bell ?
MOT:	where's the bell ?

Table 3.6: Example of Rephrasing Question

3.2.4 Feedback on Question Answer

In this category, the caregiver provides feedback on a question-and-answer sequence initiated by the child. Typically, this feedback occurs when the child answers their own question. The caregiver's role is to confirm the correctness of the child's utterance. The adult's confirmation reinforces the child's answer and potentially fosters their confidence in exploring and answering their own questions. An example is highlighted on table 3.7.

CHI:	what's that ?
CHI:	that looks like a gate .
MOT:	it does look like a gate , you're right

Table 3.7: Example of Feedback on Question Answer

3.3 Feedback Detection

To address the first research question, a two-stage approach for detecting feedback is proposed. First, a baseline system is implemented using rule-based heuristics. The details of these rules are presented in Section 3.3.1. For validation purposes, annotations for a subset of the defined feedback categories are obtained from linguistic experts. By comparing these annotations to the system's detection, the accuracy of the initial rules can then be calculated.

Building upon the evaluation of the rule-based system, a better algorithm will be developed which incorporates Large Language Models (LLMs) properties like perplexity and cosine similarity. This integration of statistical data beyond the limitations of the initial rules is expected to improve the accuracy of feedback detection. This strategy is described in Section 3.3.2.1.

3.3.1 Rule-based Feedback Detection

For the rule-based algorithm, a set of rules similar to the methodology of Hiller and Fernández (2016) was adopted. To identify patterns of repetition across the dataset, Vocabulary Overlap (VO) was used as the primary heuristic with the exclusion of stop words.

Stop words. The following is list of stop words used. It is a compilation outlined by Hiller and Fernández (2016), representing the 100 most commonly occurring words within the CHILDES dataset: ['a', 'about', 'and', 'at', 'because', 'big', 'but', 'down', 'for', 'good', 'he', 'her', 'here', 'his', 'I', 'if', 'in', 'is', 'it', 'just', 'me', 'my', 'no', 'not', 'now', 'of', 'oh', 'okay', 'on', 'out', 'right', 's', 'she', 'so', 't', 'that', 'the', 'them', 'then', 'there', 'they', 'this', 'to', 'too', 'up', 'we', 'well', 'with', 'yeah', 'yes', 'you', 'your'].

Vocabulary Overlap. To evaluate the vocabulary overlap, VO between two utterances, while excluding stop words and punctuation, equation (3.1) is used :

$$\mathrm{VO} = \frac{|A \cap C|}{|A|} \tag{3.1}$$

where: A represents the set of words in the adult utterance. C represents the set of words in the child utterance. $|A \cap C|$ represents the cardinality of the intersection between A and C. $|A|$ represents the cardinality of the set A.

3.3.1.1 Modelling Question Answer

Modelling Question Answer involves two adjacent utterances where both speakers are the same adult (a1, a2). The first utterance (a1) ends with a question mark, signifying a question posed by the adult. The second utterance (a2) is the answer provided by the same adult, lacking a question mark at the end. Algorithm 1 outlines the processing of isolating utterance pairs for this category.

Algorithm 1 Algorithm for Extracting Modelling Question Answer Pairs

```
Require: Dataframe of utterances with speaker roles
  for each utterance pair (a1, a2) in the dataframe do
      if neither speaker of (a1, a2) has the role "Target - Child" then
          if speaker of a1 is an adult and speaker of a2 is the same adult then
              if a1 ends with a question mark then
                  if a2 does not end with a question mark then
                      Store (a1, a2) as a Modelling Q-A sequence feedback pair
                  end if
              end if
          end if
      end if
  end for
```

3.3.1.2 Exemplifying Question Answer

This involves three adjacent utterances where the first utterance is a question initiated by the child (c1). The second utterance, (a1) is spoken by the adult and rephrases or repeats the child's question (c1) partially or completely. Both utterances c1 and a1 have a vocabulary overlap 0 < VO < 1. The third utterance (a2) is the answer provided by the adult, addressing the child's original question (c1). Algorithm 2 shows the process of extracting the utterance triples for this feedback category.

Algorithm 2 Algorithm for Extracting Exemplifying Question Answer Triples

Require: Dataframe of utterances with speaker roles
  for each sequence of three utterances $(c1, a_1, a_2)$ in the dataframe **do**
    if speaker c_1 has the role "Target - Child" **then**
      if speakers of (a_1, a_2) don't have the role "Target - Child" and are the same adults **then**
        if c_1 ends with a question mark **then**
          if a_1 ends with a question mark **then**
            Calculate vocabulary overlap VO between (c_1, a_1)
            if $0 < VO < 1$ **then**
              if a_2 does not end with a question mark **then**
                Store (c_1, a_1, a_2) as a exemplifying Q-A congruence feedback pair
              end if
            end if
          end if
        end if
      end if
    end if
  end for

3.3.1.3 Rephrasing Question

The first utterance (c1) is the child's question that ends with a question mark and the adult (a1) responds by rephrasing or paraphrasing the child's question (c1) entirely or partially. The adult's utterance (a1) also ends with a question mark. Algorithm 3 highlights the extraction process for utterance pairs.

3.3.1.4 Feedback on Question Answer

Feedback on Question Answer involves three adjacent utterances where the first utterance (c1) is the child asking a question. The second utterance (c2) is the child answering their own question and the third utterance (a1) is the adult's response, confirming the child's answer. Both utterances a1 and c2 have a vocabulary overlap 0 < VO < 1. Algorithm 4 outlines the extraction of the defined utterance triples.

Algorithm 3 Algorithm for Extracting Rephrasing Question Pairs

Require: Dataframe of utterances with speaker roles
  for each utterance pair (c_1, a_1) in the dataframe **do**
    if speaker c_1 has the role "Target - Child" **then**
      if speaker of a_1 does not have the role "Target - Child" **then**
        if c_1 ends with a question mark **then**
          if a_1 ends with a question mark **then**
            Calculate vocabulary overlap VO between (c_1, a_1)
            if $0 < VO < 1$ **then**
              Store (c_1, a_1) as a rephrasing child q's pair
            end if
          end if
        end if
      end if
    end if
  end for

Algorithm 4 Algorithm for Extracting Feedback on Question Answer Triples

Require: Dataframe of utterances with speaker roles
  for each sequence of three utterances $(c1, c_2, a_1)$ in the dataframe **do**
    if speakers (c_1, c_2) have the role "Target - Child" **then**
      if speaker of a_1 does not have the role "Target - Child" **then**
        if c_1 ends with a question mark **then**
          if c_2 does not end with a question mark **then**
            Calculate vocabulary overlap VO between (c_2, a_1)
            if $0 < VO < 1$ **then**
              Store (c_1, c_2, a_1) as a feedback on Q-A Dialogue triple
            end if
          end if
        end if
      end if
    end if
  end for

3.3.2 Automatic Feedback Detection

To enhance accuracy, the annotated subset of the data is used to develop a classifier capable of automatically detecting feedback within in a given dataset. The primary objective is to devise a simple and interpretable predictive model for detecting feedback on a relatively small dataset. For this task, features such LLMs perplexity, word embeddings and cosine similarity are extracted from the data and employed as features for training the predictive model. This approach seeks to evaluate whether these features can provide a more robust and reliable heuristic for detecting feedback. The subsequent sections will delineate the features and approach employed for each feedback category.

3.3.2.1 Perplexity

In LLMs, perplexity metric measures how surprised or perplexed a language model is by a given sequence of text, providing an estimate of the likelihood the model assigns to that text based on its training data (Bengio et al., 2000). Regions with higher perplexity indicate lower probability sequences that are more unexpected to the model. To explore the proposed approach, the perplexity values for each utterance in the annotated dataset for each feedback category are computed. The hypothesis is that regions of high perplexity in the dialogues will align with instances where feedback should be provided. Specifically, utterances made by children will exhibit higher perplexity compared to adult utterances, suggesting that these are the regions where feedback is required for the children. For this analysis, the pre-trained GPT-2 model (Radford et al., 2019) from the Hugging Face Transformers library (Wolf et al., 2019) is used to compute the perplexities for the utterances. For each utterance, 6 previous utterances before it, referred to as the context utterances are concatenated with the current utterance to form the input sequence. This sequence is then passed through the model to obtain the perplexity score. Listing 3.1 shows how these values are calculated.

```python
# Function to calculate perplexity score
def calculate_perplexity(utterance, context):
    # Tokenize the context and utterance
    tokenized_input = tokenizer.encode(context + " " + utterance,
    return_tensors='pt').to(device)
    # Get the perplexity score from the model
    with torch.no_grad():
        outputs = model(tokenized_input, labels=tokenized_input)
        loss = outputs.loss
        perplexity = torch.exp(loss)
    return perplexity.item()

```

Listing 3.1: Perplexity score computation

3.3.2.2 Word Embeddings

Subsequently, word embeddings, dense vector representations of words in a high-dimensional space, were calculated. Word embeddings are numerical representations that capture the semantic meaning of words (Voita, 2023). In this context, each utterance is represented as a vector based on its word embeddings. This process involved training a Word2Vec model on tokenized utterances from the dataset. Word2Vec is a popular word embedding technique developed by Mikolov et al. (2013), which learns distributed representations of words based on their co-occurrence patterns in a large corpus of text. The resulting embeddings provided a numerical representation of words, allowing for semantic similarity comparisons between utterances. Listing 3.2 shows how these values were computed.

```python
# Word embeddings
def calculate_word_embeddings(df, tokenized_column='tokenized_utterance',
     vector_size=300, window=5, min_count=1, workers=4):
    # Train Word2Vec model
    model = Word2Vec(sentences=df[tokenized_column], vector_size=vector_size, window=window, min_count=min_count, workers=workers)

    # Generate word embeddings dictionary
    word_embeddings = {word: model.wv[word] for word in model.wv.key_to_index}

    # Calculate utterance embedding
    def get_utterance_embedding(tokenized_utterance):
        embeddings = [word_embeddings.get(word, np.zeros(vector_size)) for word in tokenized_utterance]
        return np.mean(embeddings, axis=0)

    # Calculate utterance embeddings and assign to a new column in the dataframe
    df['utterance_embedding'] = df[tokenized_column].apply(get_utterance_embedding)
    return df
```

Listing 3.2: Word embeddings computation

3.3.2.3 Cosine Similarity

Utilizing the embeddings obtained, cosine similarity is calculated to quantify the semantic similarity between pairs of utterances. Cosine similarity, a metric used in NLP, measures the similarity between two vectors by calculating the cosine of the angle between them (Han et al., 2012). It is used to determine the semantic relationship between words, phrases, or documents encoded as vectors in an embedding space. In this case, the cosine similarity compared the embeddings of two utterances within the dataset. By calculating the cosine similarity between the embedding vectors of corresponding utterances in a pair, a similarity score was derived. This score indicated the degree of closeness in meaning between two utterances. Values range between -1 and 1, where a value closer to 1 indicates a higher degree of similarity between the utterances, while values closer to -1 implied dissimilarity. A score of 0 suggests orthogonality or no semantic relationship between the vectors. Listing 3.3 shows how these values were calculated.

```
# Calculate cosine similarities
def calculate_cosine_similarity_scores(df, embedding_column='
    utterance_embedding'):
    cosine_similarity_scores = []

    # Iterate through the dataframe in pairs
    for i in range(0, len(df), 2):
        # Extract embeddings for the two adjacent utterances
        embedding_1 = np.array(df[embedding_column].iloc[i])
        embedding_2 = np.array(df[embedding_column].iloc[i + 1])

        # Calculate cosine similarity between the embeddings
        similarity_score = cosine_similarity([embedding_1], [embedding_2
    ])[0][0]

        cosine_similarity_scores.extend([similarity_score,
    similarity_score])

    # Assign the similarity scores to the dataframe
    df['cosine_sim'] = cosine_similarity_scores
    return df
```

Listing 3.3: Computation of cosine similarity

3.3.2.4 Logistic Regression Classifier

Logistic Regression is chosen for this task due to its simplicity, interpretability, and effectiveness in binary classification problems (Jurafsky and Martin, 2009). Logistic regression is a widely used statistical method that models the probability of a binary outcome using a logistic function (Murphy, 2012), making it well-suited for predicting the presence or absence of feedback based on the extracted features. The decision to use this model is also due to it's robustness in handling linearly separable classes (Ng and Jordan, 2001) and its ability to provide insights into the relative importance of different features through the examination of model coefficients (Jurafsky and Martin, 2009).

To ensure balanced class representation and mitigate potential biases, equal samples of correct and incorrect feedback instances are included for each feedback category during the training process. The data for each feedback category is split into training and testing sets using `train_test_split()` method from the scikit-learn library(Pedregosa et al., 2011), with a random state of 42 to ensure reproducibility. The test set is also stratified to ensure a balanced representation of both class labels in during testing. Additionally, before training the classifier, feature scaling is applied to the training data using `StandardScaler` (Pedregosa et al., 2011). This step ensured that the features were on a consistent scale, which can improve the performance and convergence of the model.The testing data was also scaled using the same `StandardScaler` instance fitted on the training data.

The specific features used to train the classifier for each feedback category is presented in the following sections.

3.3.2.4.1 Modelling Question Answer: As detailed in 3.3.1.1, this category contains two adult utterances `(a1,a2)` in a pair. The following features were used to train the classifier:

- `adult1_perp`: This feature represents the perplexity of the first adult utterance `a1` in the pair.
- `adult2_perp`: This represents the perplexity of the second adult utterance `a2` in the pair.
- `cosine_sim`: This represents the computed cosine similarity between the vector representations of the two utterances in the pair.
- `perp_diff`: This feature is the difference between the perplexity scores of the two utterances in the pair. It helps capture the relative difference in how well the language model predicts each utterance.
- `VO`: Vocabulary overlap between the two utterances `(a1,a2)`.

3.3.2.4.2 Exemplifying Question Answer: Previously discussed in 3.3.1.2, this category contains three utterances `(c1,a1,a2)` in a triple. The following features were used to train the model for predicting this case of feedback:

- `child_perp`: This feature represents the perplexity of the child's utterance `c1` in the triple.
- `adult1_perp`: The perplexity of the first adult utterance `a1` in the triple.
- `adult2_perp`: The perplexity of the second adult utterance `a2` in the triple.
- `cosine_sim1`: This represents the computed cosine similarity between the vector representations of `c1` and `a1`.
- `cosine_sim2`: This represents the computed cosine similarity between the vector representations of `a1` and `a2`.

- `VO1`: Vocabulary overlap between the two utterances `(c1,a1)`.

- `VO2`: Vocabulary overlap between the two utterances `(a1,a2)`.

3.3.2.4.3 Rephrasing Questions: This category contains two utterances `(c1,a1)` in a pair. The following features were used to train the Logistic Regression model:

- `child_perp`: This feature represents the perplexity of the child's utterance `c1` in the pair.

- `adult_perp`: The perplexity of the adult utterance `a1`.

- `cosine_sim`: This represents the computed cosine similarity between the vector representations of `c1` and `a1`.

- `perp_diff`: This feature is the difference between the perplexity scores `child_perp` and `adult_perp`.

- `VO`: Vocabulary overlap between the two utterances `(c1,a1)`.

3.3.2.4.4 Feedback on Question Answer: With three utterance `(c1,c2,a1)` in a triple, the model was trained with the following features:

- `child_perp1`: This feature represents the perplexity of the child's first utterance `c1` in the triple.

- `child_perp2`: The perplexity of the second child utterance `c2` in the triple.

- `adult_perp`: The perplexity of the adult utterance `a1` in the triple.

- `cosine_sim1`: This represents the computed cosine similarity between the vector representations of `c1` and `c2`.

- `cosine_sim2`: This represents the computed cosine similarity between the vector representations of `c2` and `a1`.

- `VO1`: Vocabulary overlap between the two utterances `(c1,c2)`.

- `VO2`: Vocabulary overlap between the two utterances `(c2,a1)`.

3.4 Feedback Generation

This section endeavours to answer the second research question, which involves generating contextually appropriate feedback responses for children in child-adult interactions. Building upon the groundwork laid by the first research question of detecting feedback, the project now aims to generate feedback. By leveraging LLMs, the goal is to mimic the customised feedback caregivers provide to children. The annotated subset of the data will be used to generate and evaluate responses. The following sections will provide details of how this experiment is conducted.

3.4.1 Model Selection

For this study, three distinct language models will be examined to analyze the generation of contextually relevant responses in dialogues between children and adults: Open Pretrained Transformer (OPT) (Zhang et al., 2022), Dialogue Pretrained Transformer (DialoGPT) (Zhang et al., 2019b) and Gemma (Gemma et al., 2024).

- **Open Pretrained Transformer**: OPT was developed by Meta AI, trained on a vast corpus of online data using self-supervised learning techniques (Zhang et al., 2022). OPT serves as a strong baseline model due to its impressive generative capabilities and ability to handle diverse language tasks (Zhang et al., 2022). It will function as the primary model for exploring the effectiveness of a general-purpose large language model in generating feedback responses within child-adult dialogues.

- **Dialogue Pretrained Transformer**: DialoGPT is a dialogue-specific language model developed by Microsoft (Zhang et al., 2019b). It is pre-trained on a massive corpus of conversational data, specifically optimized for dialogue understanding and generation tasks. DialoGPT's architecture is designed to capture the nuances of multi-turn conversations, positioning it as a highly promising candidate for the precise task of providing relevant responses in multil turn dialogues between children and caregivers.

- **Gemma**: Gemma, a recent LLM released by Google in February 2024 (Gemma et al., 2024) is available in two different sizes. The first is a 7 billion parameter model designed for efficient deployment and development on GPU and TPU. The second is a 2 billion parameter model tailored for CPU and on-device applications. Gemma 2B and 7B have been trained on 2T and 6T tokens respectively, drawing primarily from English data extracted from web documents, mathematical sources, and code (Gemma et al., 2024). For this experiment, the 2B base version of the model will be used. The following key features of Gemma where considered for the selection:

- **Training Data Filtering** - The training data undergoes meticulous filtering to remove sensitive information and potentially unsafe content like Child Sexual Abuse Material (CSAM), hate speech and violent content (Gemma et al., 2024).
- **Lightweight** - Compared to other open models, Gemma achieves strong performance while keeping its size relatively small which allows it to run on devices with less processing power (Gemma et al., 2024).
- **Reinforcement Learning Integration** - Gemma's fine-tuning capabilities extend to reinforcement learning (RL). This approach involves training the model through feedback mechanisms, where it learns by receiving rewards for desired outputs and penalties for undesired ones, particularly useful for tasks where human evaluation of outputs plays a crucial role (Gemma et al., 2024).

One critical aspect of this thesis involves ensuring that the feedback generated adheres to safety standards and is age-appropriate for children. The selection of Gemma was influenced by the strong dedication of its developers to safety protocols throughout the training phase (Gemma et al., 2024).

Moreover, the architectural design of the model promotes responses that are suited to the specific context, rendering it highly appropriate for the given objective. Gemma employs an architecture based on a transformer-decoder model, enabling thorough analysis of the entire dialogue history prior to response generation. This emphasis on contextual comprehension is critical for producing feedback that is useful for children, consequently, it is expected that Gemma will outperform the other two models.

3.4.2 Data Pre-processing for Fine-tuning

This section outlines the key steps undertaken to prepare the raw data for generation. These steps aim to improve data quality and ensure interpretability for the models.

- **Isolation Annotated Dialogues:** The pre-processing began by identifying and removing all dialogues that were in the subset of data that was annotated. This is to prevent training on the test set as the annotated data will be employed for testing.

- **Unintelligible Utterance Removal:** The next step was identifying and removing unintelligible or uninterpretable variables that wouldn't contribute meaningfully to the analysis. Instances of utterances containing these were hand analysed. The manual inspection aimed to identify and exclude any utterances deemed unintelligible or irrelevant as described in MacWhinney (2000). The variables removed include : [xxx, xx, yyy, @, <>,:, 0, &, &-].

- **Censoring Names:** Some names within the data were anonymized to protect privacy. Nouns containing 'www' were assumed to be censored names. A dictionary of stand-in names form Kantrowitz (1994) was used to replace censored names while maintaining some level of naturalness.

- **Concatenating Utterances:** Consecutive utterances from the same speaker were concatenated into speaker turns, providing a more coherent representation of the dialogues. Each utterance was tagged with a speaker identifier, where 'A' denotes an adult utterance, and 'C' represents a child utterance. Subsequently, the utterances were formatted into a text file, transitioning from their original dataframe structure to a format suitable for further analysis.

- **Train-validations Split:** For this study, 20% of the data was spilt for validation while the remaining 80% was used as the train set.

Table 3.8 presents a summary of the data used for both training and validation, including the total number of dialogues and utterances after the pre-processing steps. The complete pre-processing pipeline, encompassing all the transformations applied to the raw data, is detailed in Appendix A.

	Dialogues	Adult Utterances	Child Utterances
Train	1,436	781,287	364,246
Validation	359	107,315	94,446
Total	1,795	888,602	458,692

Table 3.8: Overview of dataset for training and validation

3.4.3 Model Fine tuning

3.4.3.1 OPT and Dialogpt

The fine-tuning process was carried out using the Hugging Face Transformers library (Wolf et al., 2019) in Python. The 350 million parameter version of the OPT model while the medium-sized version of the DialoGPT was used. The two models are fine-tuned for 5 epochs, using an early stopping to save the best performing model based on its evaluation loss.

3.4.3.2 Gemma

For fine-tuning Gemma, the QLoRA technique introduced by Dettmers et al. (2023), which combines Quantization and Low Rank Adaptation (LoRa) was used. Quantization is a technique used to reduce memory usage and inference time of deep learning models by converting the model's weights and activation from high-precision data types to lower-precision data types.

LoRA differs from traditional fine-tuning methods by freezing pre-trained model weights and introducing trainable rank decomposition matrices, significantly reducing the number of trainable parameters (Hu et al., 2021).

QLoRA optimizes memory usage without compromising performance, such as the use of a 4-bit NormalFloat (NF4) data type, Double Quantization to reduce memory footprint, and Paged Optimizers to manage memory spikes (Dettmers et al., 2023). The method has been shown to outperform previous models on the Vicuna benchmark and enables the fine-tuning of large language models on a single GPU (Dettmers et al., 2023), making it a significant advancement in the field of Natural Language Processing.

For this task, the Gemma 2B model, a 2 billion parameter version of the model, was quantized. The quantization process was configured using the BitsAndBytesConfig class from the Bitsandbytes library (Dettmers et al., 2023). The 4-bit NormalFloat (nF4) quantization was used and a LoRa value of 8 is utilized for computational efficiency during training.

The following parameters were set:

- `load_in_4bit = True:` This option enabled loading the model weights in a 4-bit quantized format, thereby reducing the memory requirements by a factor of 8 compared to the default 32-bit setting.
- `bnb_4bit_quant_type= nf4:` The 'nf4' quantization type was used, which means NormalFloat 4-bit quantization. This quantization method aims to minimize the quantization error by adaptively determining the quantization ranges based on the weight distribution.
- `bnb_4bit_compute_dtype = torch.bfloat16:` During the forward and backward passes, the quantized weights were temporarily de-quantized to bfloat16 (brain floating-point 16) format for computations. This approach strikes a balance between precision and computational efficiency, as bfloat16 offers higher precision than traditional float16 while still providing significant speed-ups compared to float32 computations.

Other than the specific QLoRa parameters tailored for Gemma, the training settings applied to OPT and DialoGPT remained consistent. The model is fine-tuned on Google Colab Pro for 2 hours using the A100 GPU runtime.

3.4.4 Model Generation Procedure

The annotated datasets for the Rephrasing Questions and Feedback on Question Answer categories were used as test sets for evaluating the models' performance. For each feedback instance in the dataset, a prompt was constructed by providing six utterances as context, with the speaker label ('A' for adult, 'C' for child) appended to each utterance. The final speaker label in the prompt corresponded to the adult target feedback that the model was tasked with generating. To account for the stochastic nature of language generation and capture the diversity of potential responses, five candidate utterances were generated per prompt.

The format of the prompt used for testing is illustrated in Figure 3.1, where each prompt consists of the context utterances and the corresponding speaker labels, followed by speaker label of the target text. Text highlighted in red indicates the utterance that requires feedback. Speaker label marked in blue denotes text to be generated.

The specific parameters employed during the generation process are detailed in Subsection 3.4.4.1.

Prompt
A: who is that ?
C: Gordon .
A: what's he doing ?
C: sitting on the couch .
A: sure he is .
C: what i Fraser doing ?
A:

Figure 3.1: Example prompt used as input for text generation

3.4.4.1 Generation Parameters

The generation parameters were chosen to strike an appropriate balance between output quality, diversity, and relevance for the child-directed dialogue context being evaluated. The following parameters were used to generate responses from the fine-tuned models:

- **Maximum New Tokens:** The maximum length of the token was set to 10. This length was chosen based on an analysis of the child-adult dialogue corpus, which showed that typical feedback or response utterances from adults tended to be relatively short, averaging around 7-8 words. This is to ensure that the generated outputs were of a suitable length for providing feedback responses.

- **Number of Return Sequences:** For each prompt, the models generated 5 candidate responses. This allowed for a diverse set of potential responses to be evaluated, rather than relying on just the single top-generated output (Molnar et al., 2023).

- **Top-k Sampling:** Top-k sampling was used, where the token probabilities were truncated to only consider the top 50 most likely tokens during the generation process. A top-k value of 50 was selected as it has been shown to strike a balance between diversity and quality in text generation tasks (Holtzman et al., 2019). Using a smaller top-k value (e.g., 10) can lead to more repetitive and less diverse outputs, while a larger top-k (e.g., 100) may introduce more noise and lower-quality tokens. The choice of 50 was intended to encourage the models to explore a wider range of potentially relevant tokens while still maintaining a reasonable level of quality in the generated text.

- **Re-ranking:** The generated top-k outputs are then re-ranked based on their vocabulary overlap with the erroneous child utterance that requires feedback. This re-ranking step prioritized generated utterances that a more directly relevant as feedback to the child's input, since the focus of the evaluation was on the models' ability to produce appropriate feedback.

Prior to generation with the fine-tuned models, a baseline generation was conducted using the settings outlined above. This baseline generation allows for a direct comparison with the fine-tuned versions, enabling a better analysis of the impact of domain-specific fine-tuning on the models' performance.

3.4.5 Evaluation

The performance of the three models are assessed through a combination of automatic and human evaluation. Subsections 3.4.5.1 and 3.4.5.2 detail the metrics used for the automatic evaluation and how the human evaluation was conducted respectively.

3.4.5.1 Automatic Evaluation

Each model's five generated sequences undergo comparison with the gold reference text. Subsequently, the mean value is calculated across the two feedback categories used for testing and across all three models. For this assessment, the following metrics are utilized:

- **BLEU**: BLEU, a widely used metric, measures the n-gram overlap between the generated text and the reference text (Papineni et al., 2002). This will indicate how natural and human-like the generated language is compared to real adult feedback responses. While BLEU is utilized as an evaluation metric in this study, it is important to note that it is not the primary metric, due to concerns about its validity and reliability, including varying correlations with human evaluations based on contextual factors (Reiter, 2018). It is included as an additional evaluation measure to provide a quantitative assessment of language generation quality alongside other metrics.

- **BERTScore**: Additionally, the BERTScore metric (Zhang et al., 2019a) has been employed to evaluate the semantic similarity between the generated text and reference text, leveraging the contextualized representations learned by BERT (Devlin et al., 2018). This metric can capture the coherence and appropriateness of the generated language beyond simple lexical overlap. BERTScore provides stronger model selection performance than existing metrics, indicating its effectiveness in evaluating generated texts.

- **Mauve Score**: MAUVE compares machine-generated text to human-written text by directly comparing the patterns in text created by machines to those in text written by humans (Pillutla et al., 2021). MAUVE offers a comprehensive and reliable way to measure how closely the machine-generated text aligns with human-written text. MAUVE captures a wide range of errors and nuances present in both types of text, providing a more holistic assessment. Additionally, MAUVE has been shown to correlate well with human judgments and can identify quality differences based on various factors like text length, decoding algorithm, and model size (Pillutla et al., 2021). By using MAUVE, a more accurate evaluation of the generated texts can be achieved.

3.4.5.2 Human Evaluation

While automatic evaluation metrics provide a useful quantitative assessment of the models performance in the generation task, human evaluation remains a critical component for assessing the quality and appropriateness of the generated text (Belz and Reiter, 2006), especially in the context of child-directed interactions. To ensure a more effective human evaluation process, this research follows several of the best practices and recommendations outlined by Van Der Lee et al. (2019) for the evaluation of automatically generated. The human evaluation procedure is outlined below:

3.4.5.2.1 Consent Form: Before participation in the human evaluation, all participants are required to read and agree to a detailed consent form outlining the purpose and procedures of the research study. The consent form emphasizes that participation was voluntary, and participants could withdraw at any time without providing a reason, up until the point of analysis. The consent form informs participants that the information and data they provided would be securely stored. Participants are asked to confirm their understanding and acceptance of these data usage and sharing practices. By requiring participants to thoroughly review and provide consent for their participation, the study ensured that human raters are fully informed about the research project and their role, while also safeguarding their rights and privacy.

3.4.5.2.2 Evaluation Criteria: In line with Van Der Lee et al. (2019), who emphasized the importance of using separate criteria instead of relying on an overall quality assessment when evaluating language generation systems, a set of evaluation criteria are employed to capture multiple dimensions of the generated feedback. The following criteria are selected based on their relevance to the context of child-directed speech:

- **Fluency:** Evaluating if the generated feedback is grammatically correct and fluent in English, regardless of context. This criterion is commonly used in evaluating language generation systems, as fluency and grammaticality are fundamental requirements for natural-sounding text (Novikova et al., 2017).

- **Coherence:** Assessing if the generated feedback aligns with the content and context of the original utterance, making sense given the conversational flow. Coherence is critical in dialogue systems, as generated responses should be contextually appropriate and maintain the logical flow of the conversation (Dziri et al., 2019).

- **Correctness:** Judging whether the generated feedback accurately addresses any language structure or errors present in the original child utterance. This criterion is particularly relevant in the context of child language learning, as caregivers often provide feedback to correct or rephrase children's utterances (Chouinard and Clark, 2003; Saxton, 2000).

- **Appropriateness:** Determining if the generated feedback sounds natural and suitable for the given conversational situation, considering factors like tone and age-appropriateness. The language used when interacting with children should be tailored to the child's developmental level (Snow, 1972).

By selecting these criteria, the human evaluation aims to provide a well rounded assessment of the generated feedback's quality for child-directed speech.

3.4.5.2.3 Sample Size: A sample of 28 contexts is carefully curated, 14 from each feedback category. Selected examples are representative of the two different types of feedback in the test dataset.

Distribution Across Survey - The selected examples are evenly distributed across seven separate surveys. Each survey contains four examples, with two examples from each feedback category.

Participant Evaluation - Seven participants evaluate each survey independently. This results in a total of 49 individuals contributing to the evaluation process.

While the sample size was not determined through a specific statistical formula or power analysis, the involvement of multiple annotators per survey and the inclusion of examples from both feedback categories are intentional measures to enhance the reliability and validity of the evaluation results.

3.4.5.2.4 Inter Annotator Agreement (IAA): This metric measures the level of consensus or agreement among multiple human raters or annotators when evaluating or labeling the same set of items (Artstein, 2017). The rationale for measuring agreement lies in its ability to gauge the reliability of the annotation process. A high IAA indicates a strong level of agreement, while a low IAA suggests discrepancies or inconsistency in evaluations, potentially leading to an unreliable annotation process (Artstein, 2017). Following the completion of the human evaluations, the IAA scores are then calculated to quantify the consistency of assessments among participants. However, it's important to note that while agreement is necessary, it alone does not guarantee correct annotations (Artstein, 2017).

3.4.5.2.5 Recruitment of Participants: Participants for this study are recruited from the authors' social circle, ensuring that all participants are native English speakers. Most of the participants are either care givers or have regular exposure with children. Invitations are sent out via social media platforms and email, inviting individuals to participate in the survey. The research is clearly explained to participants and they are given the chance to ask any questions. Participants are provided with a link to a Microsoft Form and invited to contribute to the completion of the survey.

3.4.5.2.6 Experiment Setup: Participants are presented with sample dialogues between a child and their caregiver, along with feedback responses from the caregiver with the age of the child clearly stated. Additionally, responses generated by the three models were provided. Participants are asked to rate the model generated responses on a 5-point Likert scale based on the evaluation criteria outlined in 3.4.5.2.2. Figures A.1 and A.2 in Appendix A show sample generations that are used for the human evaluation for both feedback categories. Utterance requiring feedback is clearly stated.

Chapter 4

Results

Building on the methods outlined in Chapter 3, this chapter presents the results of the experiments conducted to answer the research questions.

4.1 Exploratory Data Analysis

A preliminary Exploratory Data Analysis (EDA) was carried out on the curated CHILDES dataset described in Section 3.1.1 This section will present the results of the EDA to provide insights into the data. Through statistical techniques and visualizations, this analysis sought to uncover patterns and trends within the data.

Table 4.1 show the distribution of dialogues by speakers for the 2,026 dialogues present in the dataset. The results presented in Table 4.2 provide a statistical summary of the various data attributed for both children and adults. Key metrics such as Morpheme Count, Total Morphemes, Total Utterances, and Mean Length of Utterance (MLU_{word}) reveal notable differences between the two groups, indicating distinctions in language usage.

Table 4.3 shows distribution of each age group for the 46 children in the dataset. The 18-23 months age has the highest frequency of children while 24-29 months has just one child. The distribution of age groups by dialogues is shown in Figure 4.1. Majority of the dialogues have children in the 30 - 35 months age group while the 12 - 17 months age group appear the least in dialogues.

Figure 4.2 shows the relationship between a child's age in months and their MLU_{word}. The data points are scattered across the age range from 12 months to 48 months. The overall trend shows that as a child's age increases, their MLU_{word} also tends to increase. However, there is significant variation in MLU_{word} at each age, indicating individual differences in language development rates.

Number of Speakers	**Number of Dialogues**
2	544
3	982
4	402
5	98

Table 4.1: Number of Speakers and Total Number of Dialogues

	Children			Adults		
	M±Std	Med.	Max	M±Std	Med.	Max
Morpheme Count	2.43 ± 2.23	2.0	68	4.67 ± 3.66	4.0	122
Total Morphemes	1443.08 ± 1004.42	1258	8885	3426.5 ± 1766.88	3452	9277
Total Utterances	547.17 ± 292.08	534	2504	731.63 ± 359.43	763	2563
$\mathrm{MLU}_{\mathrm{word}}$	2.75 ± 0.88	2.631	7.353	4.67 ± 0.98	4.653	24
Age (months)	31.43 ± 6.82	31	48			

Table 4.2: Statistical Summary of Data Attributes

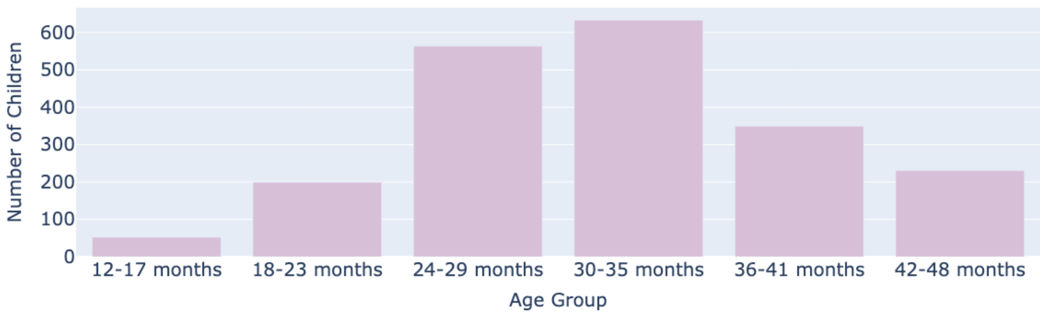

Figure 4.1: Age Group Distribution by Dialogues

Age Bin	**Number of Children**
18-23 months	16
30-35 months	13
42-48 months	9
36-41 months	4
12-17 months	3
24-29 months	1

Table 4.3: Age group of children in dataset

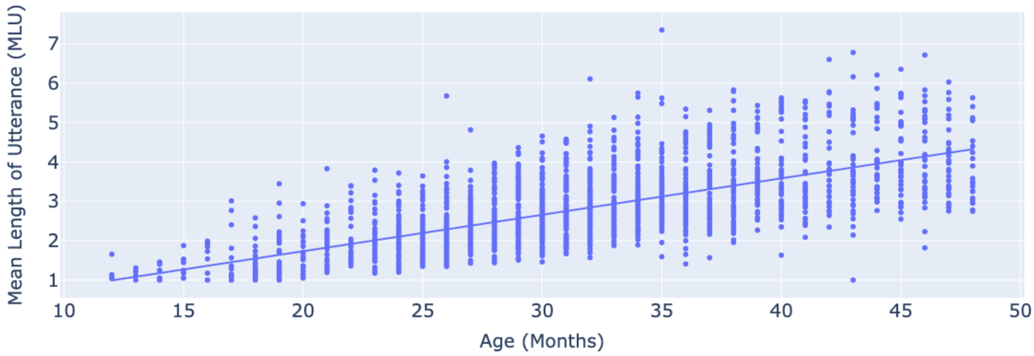

Figure 4.2: Relationship between age and MLU_{word}

4.2 Feedback Detection

This section presents the results of feedback detection using both the rule - based and automatic detection approach which have been detailed in Chapter 3. The results of both methods are presented in Sections 4.2.1 and 4.2.2 respectively.

4.2.1 Rule Based Feedback Detection

Following the application of the rule-based approach to extract different categories of feedback from the dataset, a total of **89,841** question-answer pairs and triples are isolated across the various feedback categories. After the extraction, a subset, which is representative of all the age groups in the dataset is annotated by Linguistic experts. Table 4.4 provides a detailed breakdown of these filtered sequences, presenting the distribution across each feedback category. Additionally, the table includes the total number of annotated samples within each category.

After annotations, the accuracy of the rule-based approach in detecting feedback instances across the four different categories is evaluated. The results are presented below:

- **Modelling Question Answer:** For this category, the rule based detection system achieved a correct classification rate of 68 instances out of 201, indicating an accuracy rate of approximately 33.8%.

- **Rephrasing Question:** Similar to Modelling, the rule based detection system demonstrated an accuracy rate of 34.5%, correctly identifying 69 out of 201 instances in this category.

- **Exemplifying Question Answer:** In this category, the rule based detection system demonstrated a higher accuracy level, correctly identifying 148 out of 199 instances, resulting in an accuracy rate of 74.4%.

- **Feedback on Question Answer:** The rule based detection system exhibited moderate accuracy in identifying feedback within this category, with 80 out of 190 instances correctly classified, yielding an accuracy rate of approximately 42.1%.

Overall, the rule-based detection system displayed varying degrees of accuracy across different feedback categories. While certain categories, such as Exemplifying, showed higher accuracy rates, others, like Modelling and Rephrasing , demonstrated relatively lower accuracy levels. Table 4.5 presents the accuracy of the rule based detection.

	Sequences Extracted	Sequences Annotated
Modelling	82,409	201
Rephrasing	5,540	201
Exemplifying	1,296	199
Feedback	596	190
Total	89,841	791

Table 4.4: Total number of extracted and annotated sequences for feedback categories

Feedback Type	Accuracy
Modelling	33.8%
Exemplifying	74.4%
Reprashing	34.3%
Feedback	42.1%
Overall	46.1%

Table 4.5: Accuracy of rule based feedback detection

4.2.1.1 Analysing Feedback Across Age Groups

For the set annotated as correct, Figures 4.3 and 4.4 show the count and distribution across age (in months) for the different categories. Feedback on Question Answer has the highest median age, indicating that this type of feedback is typically given at an older age compared to the other groups. Modelling and Rephrasing categories are more common among younger individuals. The Exemplifying category has a relatively wide range of ages, showing a larger variability in the ages at which this type of feedback occurs. There is a clear trend where Modelling and Rephrasing interventions decline as the child gets older, while Feedback and Exemplifying become more skewed towards older ages. This distribution aligns with the changing needs of the child during language development, where more explicit modeling and correction is required initially, but as they progress, more affirmation and examples become beneficial to support their learning.

Figure 4.3: Age Distribution per Feedback Category

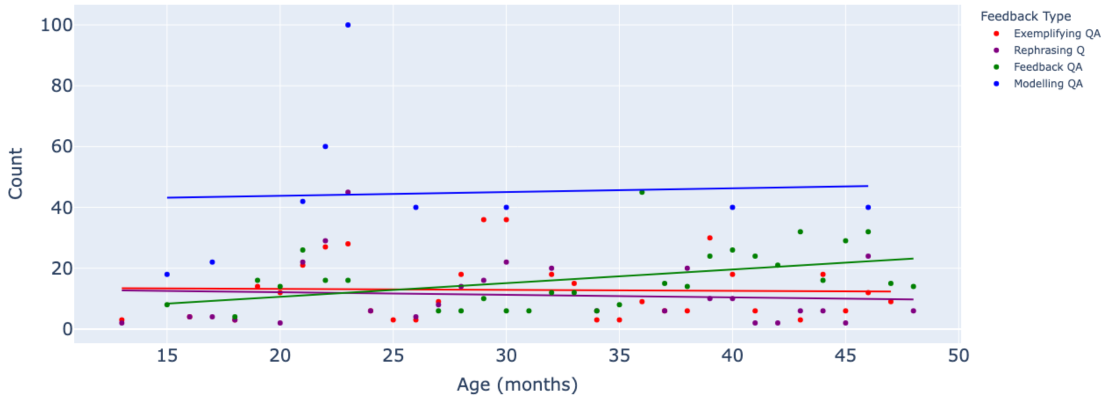

Figure 4.4: Feedback Count Across Age

4.2.2 Automatic Feedback Detection

This section will present the results of experiment carried out for automatic detection of feedback using language models properties.

4.2.2.0.1 Perplexity: For all feedback categories where both children and adults contribute to question-answer sequences, higher perplexity scores are observed for children compared to adults. The highest perplexity score for children is seen in Exemplifying Question Answer and the lowest in Rephrasing Question. Figures 4.5 and 4.6 and Table 4.6 present these results.

Additionally, as observed in Figure 4.7 perplexity of children utterances tends to decrease with increase in age suggesting that they perform better language structures as they grow older, hence LLMs are less perplexed by their utterances.

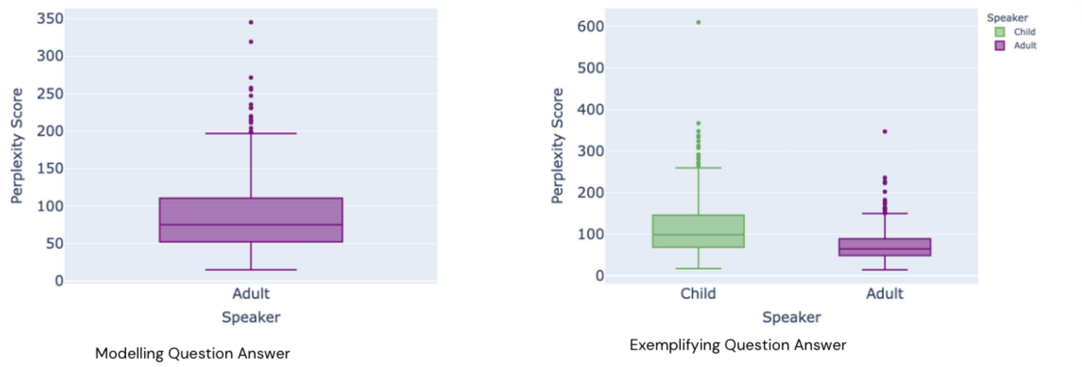

Figure 4.5: Distribution of perplexity for Modelling and Exemplifying

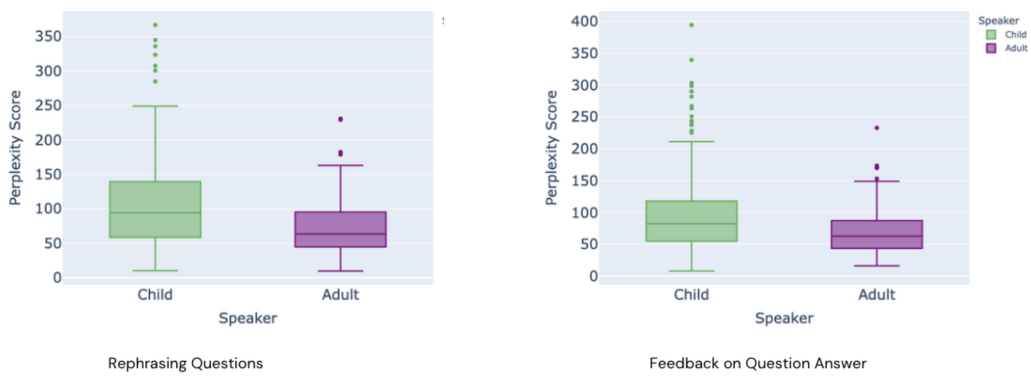

Figure 4.6: Distribution of perplexity for Rephrasing and Feedback

4.2.2.1 Logistic Regression Classifier

Table 4.7 presents the number of each class label used for the classifier for the different feedback categories. The results obtained are discussed in detail below:

4.2.2.1.1 Modelling Question Answer: The model achieved an accuracy of 61%. The F1-score for the positive class is 62%, while for the negative class, it is 59%. The full classification report is presented in Table 4.8. Out of 28 test samples, the trained classifier correctly predicted 9 correct feedback instances as correct (True Positives) and 5 correct feedback instances as

	Children			Adults		
	M±Std	Med.	Max	M±Std	Med.	Max
Modelling Q-A				87.8 ± 52.2	75.1	345.1
Exemplifying Q-A	119.6 ± 79.8	98.5	610	73.8 ± 38.9	64.8	347
Rephrasing Q	108.8 ± 68.9	94.1	366.8	72.9 ± 39.8	63.4	230.9
Feedback on Q-A	94.5 ± 56.6	82.4	394.5	69.9 ± 35.4	86.9	232.9

Table 4.6: Statistical summary of child and adult perplexity scores

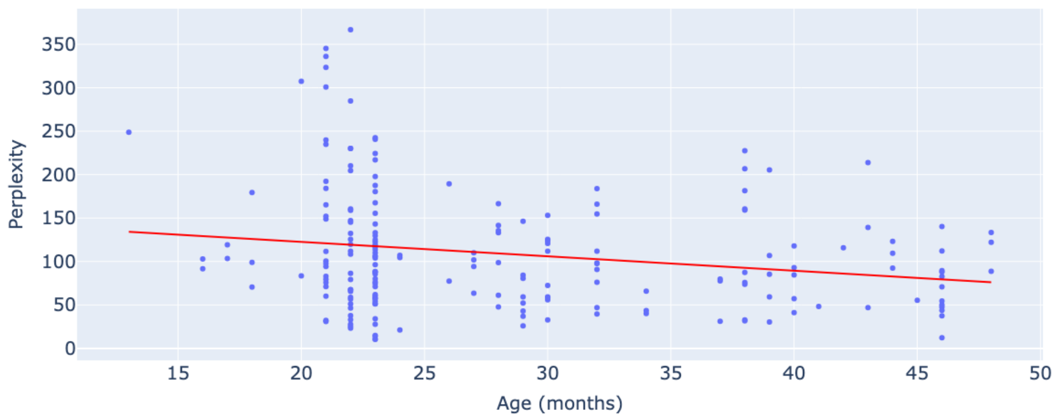

Figure 4.7: Perplexity trend with age

	Modelling	Exemplifying	Rephrasing	Feedback
correct	68	51	69	80
incorrect	68	51	69	80
Overall	136	102	138	160

Table 4.7: Total number sequences used for each feedback category classifier

incorrect (False Positive). The feature importance chart shows that the feature `adult2_perp` has the highest importance, followed by `perp_diff` and `adult1_perp`. Vocabulary overlap, `VO` and `cosine_sim` show the least importance in predicting class labels for this category. Figures 4.8 and 4.9 present these results graphically.

	precision	recall	f1-score	support
0	0.62	0.57	0.59	14
1	0.60	0.64	0.62	14
accuracy			0.61	28
macro avg	0.61	0.61	0.61	28
weighted avg	0.61	0.61	0.61	28

Table 4.8: Logistic Regression classification report for Modelling

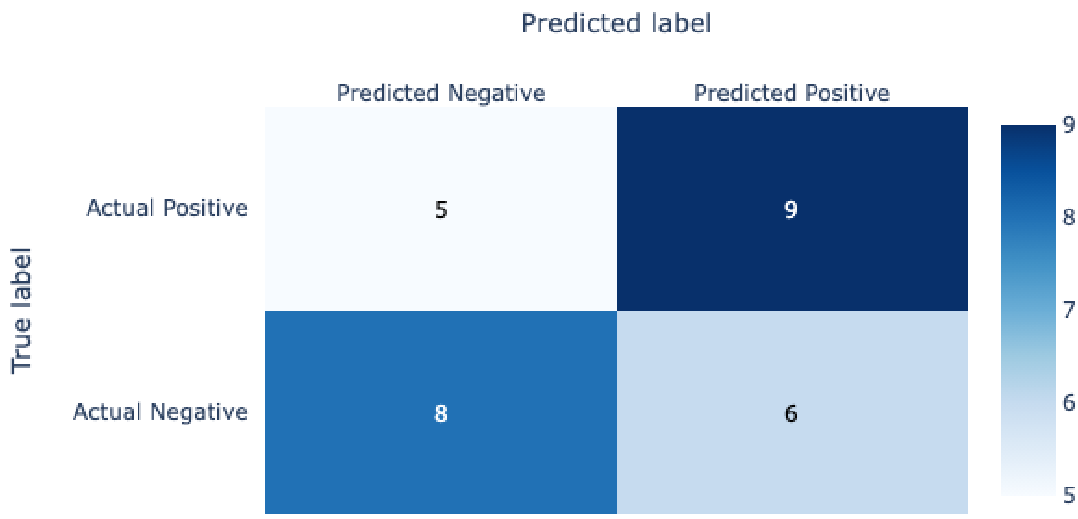

Figure 4.8: Modelling Question Answer: Confusion Matrix

4.2.2.1.2 Exemplifying Question Answer: As shown in the classification report in Table 4.9, the overall accuracy of the model is 71%. For class 0, the F1-score is 70%, and for class 1, it is 73%. Figures 4.10 and 4.11 present the confusion matrix and feature importance plot respectively. The model correctly predicted 8 correct feedback instances as correct (True Positives)

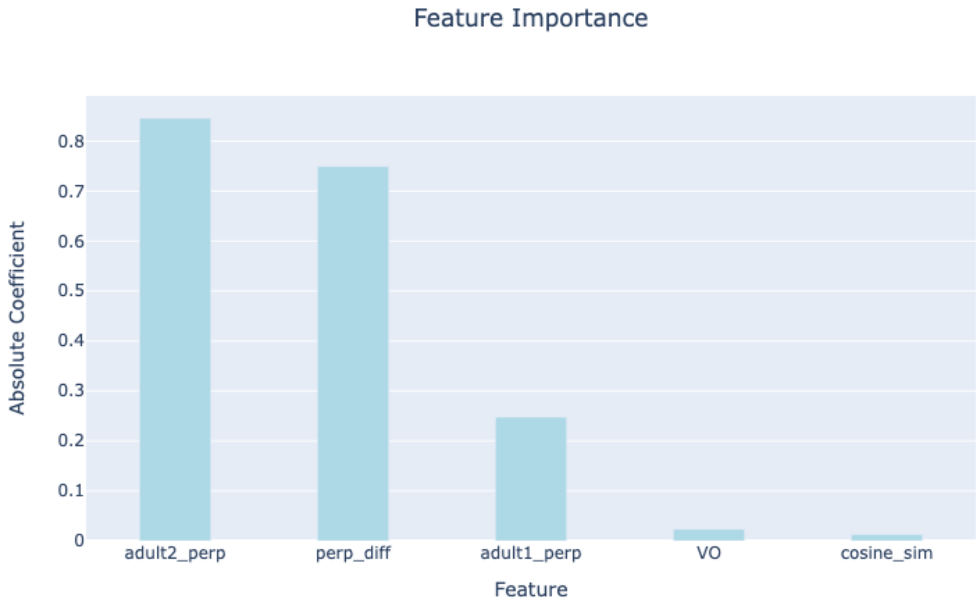

Figure 4.9: Modelling Question Answer: Feature Importance

and just 2 correct feedback instance as incorrect (False Positive). For the incorrect feedback instances, class (0), the model correctly predicted 7 as incorrect feedback instances (True Negatives) and misclassified 4 instances as correct feedback (False Positives). The most important feature guiding prediction is `adult2_perp`. `VO1` and `cosine_sim2` are the least important features guiding prediction.

	precision	recall	f1-score	support
0	0.78	0.64	0.70	11
1	0.67	0.80	0.73	10
accuracy			0.71	21
macro avg	0.72	0.72	0.71	21
weighted avg	0.72	0.71	0.71	21

Table 4.9: Logistic Regression classification report for Exemplifying

4.2.2.1.3 Rephrasing Question: Table 4.10 shows the F1-score for class 0 (incorrect feedback instances) is 58%, and for class 1, it is 62%. The overall accuracy of the model is 0.64, indicating that it correctly classified 64% of the instances. Overall, the model performs better in identifying

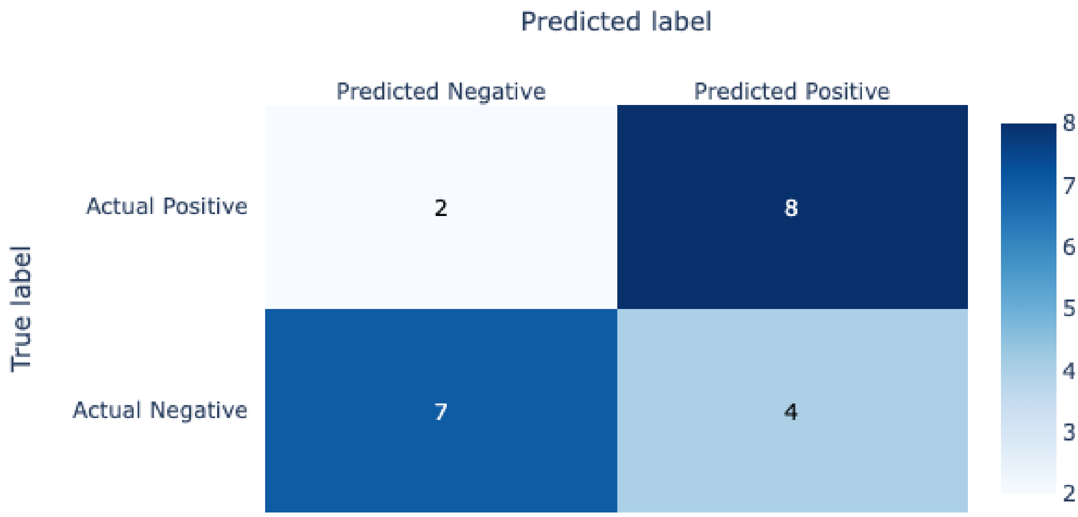

Figure 4.10: Exemplifying Question Answer: Confusion Matrix

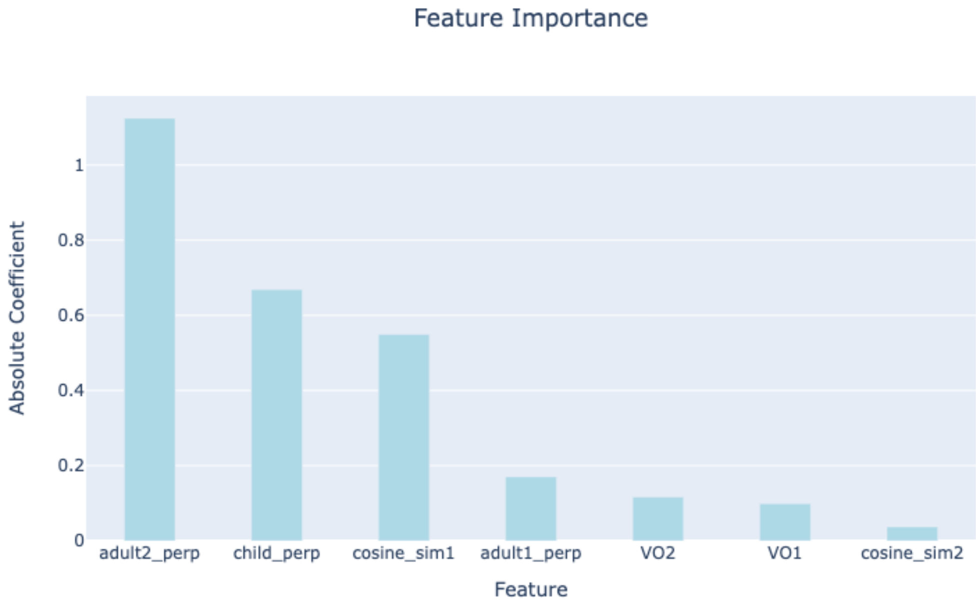

Figure 4.11: Exemplifying Question Answer: Feature Importance

the negative class (higher recall) but does better with precision for the positive class. From Figure 4.12, for the correct feedback instances, the model correctly predicted 8 instances as correct feedback and misclassified 6 instance as incorrect feedback. For the incorrect feedback class, the model correctly predicted 10 instances as negative and misclassified 4 instances as feedback. The feature importance chart displays the most important feature is `perp_diff` followed by `cosine_sim` with `VO` being the least important feature.

	precision	recall	f1-score	support
0	0.62	0.71	0.67	14
1	0.67	0.57	0.62	14
accuracy			0.64	28
macro avg	0.65	0.64	0.64	28
weighted avg	0.65	0.64	0.64	28

Table 4.10: Logistic Regression classification report for Rephrasing

4.2.2.1.4 Feedback on Question Answer: Table 4.11 presents the classification report, highlighting that the model achieved an accuracy of 84% and an an F1-score of 84%. Overall, the model seems to perform reasonably well, with a high recall for detecting positive cases and a

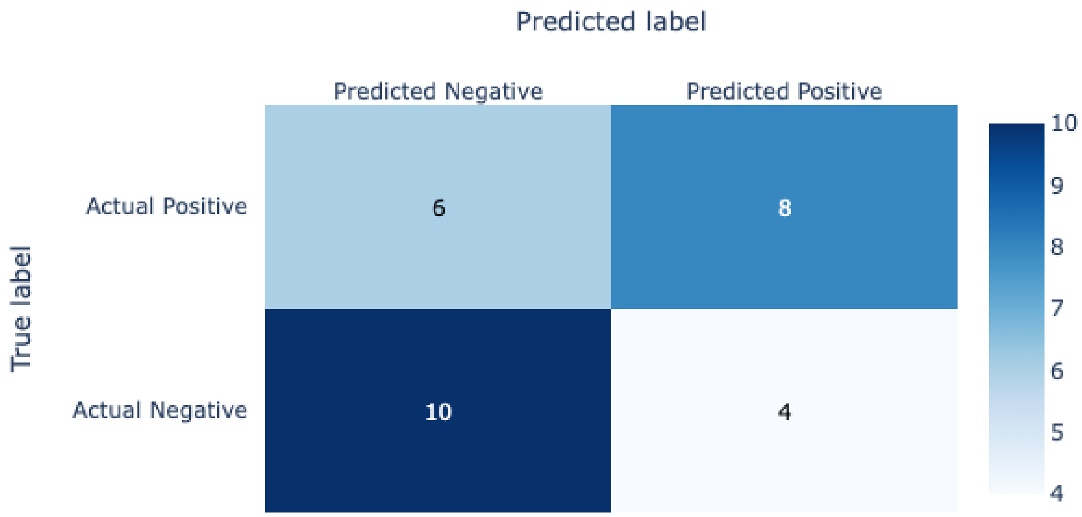

Figure 4.12: Rephrasing Question: Confusion Matrix

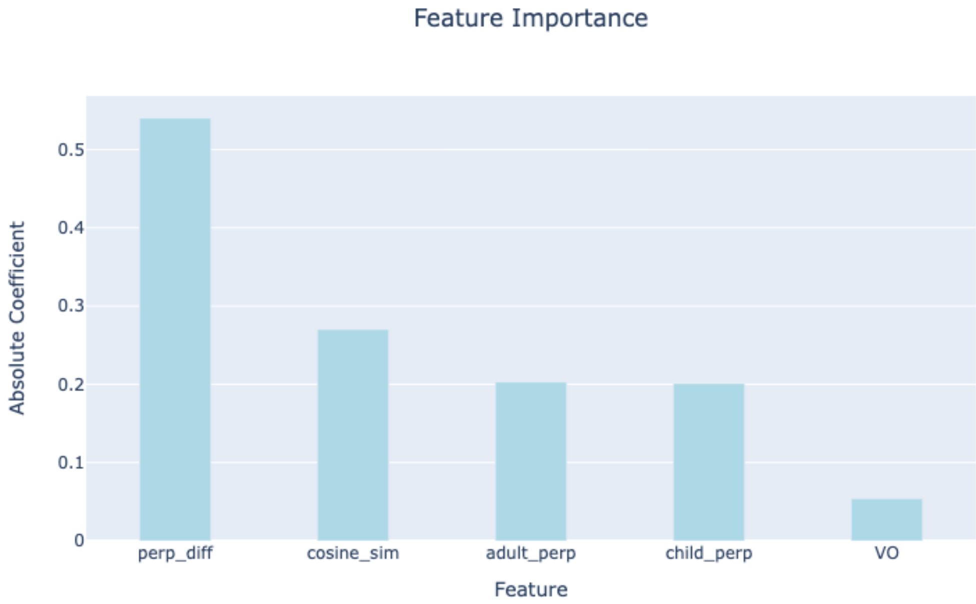

Figure 4.13: Rephrasing Question: Feature Importance

decent precision and accuracy. Figures 4.14 and 4.15 show and the confusion matrix and feature importance graph. The model correctly predicted 12 negative cases and 15 positive cases. However, it mislabeled 4 negative cases as positive (False Negative) and 1 positive case as negative (False Positive). The bar chart shows `child_perp2` and `adult_perp` had the highest importance scores, suggesting they were the most influential features for the model's decision-making process while `VO2`is the least important feature.

	precision	recall	f1-score	support
0	0.92	0.75	0.83	16
1	0.79	0.94	0.86	16
accuracy			0.84	32
macro avg	0.86	0.84	0.84	32
weighted avg	0.86	0.84	0.84	32

Table 4.11: Logistic Regression classification report for Feedback

Table 4.12 shows the overall accuracy results for the classifier in each feedback category. An increased accuracy is observed from the rule based algorithm.

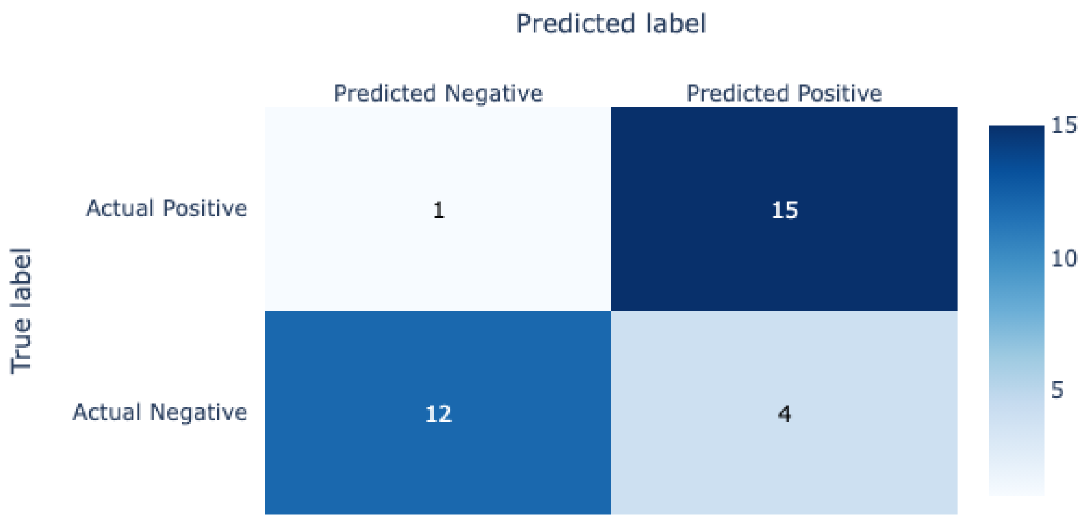

Figure 4.14: Feedback on Question Answer: Confusion Matrix

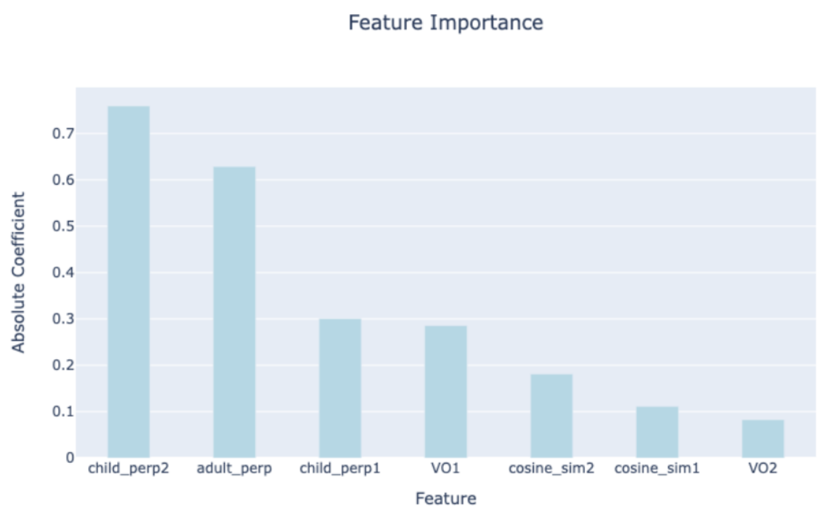

Figure 4.15: Feedback on Question Answer: Feature Importance

Feedback Type	Accuracy
Modelling	61%
Exemplifying	71%
Rephrasing	64%
Feedback	84%
Overall	70%

Table 4.12: Overall accuracy of Logistic Regression Classifier

4.3 Feedback Generation

This section presents the results of the experiment conducted on feedback generation using Language Models (LLMs), detailed in Section 3.4. By employing the fine-tuned models, the aim is the assess their effectiveness in generating contextually appropriate feedback for children. The results of both human and automatic evaluations are presented to assess the performance of each model.

4.3.1 Automatic Evaluation

The Automatic evaluation metrics described in 3.4.5.1 were used to get a quantitative measure of the performance of the models on the text generation task. Table 4.13 presents a comparison of the BLEU, BertF1 and Mauve scores of the base and fine-tuned models for Rephrasing Questions and Feedback on Question Answer categories. The fine-tuned OPT model achieved the best performance based on the reference based metrics, BLEU and BertF1 scores in both categories with values of **0.049** and **0.741** respectively while the tuned version of the DGPT model had the best Mauve score of **0.101**.

For Rephrasing Questions, the fine-tuned OPT model showed improvements from the base model, achieving a BLEU score of **0.049** and a BertF1 score of **0.733** with DGPT having the highest Mauve score of **0.021** in this category. Despite the fine-tuning efforts, there wasn't any improvement observed in the Mauve score in OPT. In fact, there was a slight decrease in performance in this regard, although the magnitude of this change was marginal. Conversely, the DGPT and Gemma model showed improved performance across all metrics following fine-tuning.

In Feedback on Question Answer, the same trend is seen with the fine-tuned OPT model achieving the best BLEU score of **0.036** and a BertF1 score of **0.741** and DGPT having the highest Mauve score of **0.101**. For this category, fine-tuning led to improved performance across all models.

		BLEU ↑	BertF1 ↑	Mauve ↑
Rephrasing				
OPT	Base	0.023	0.680	**0.007**
	Tuned	**0.049**	**0.733**	0.004
DGPT	Base	0.010	0.069	0.004
	Tuned	**0.036**	**0.443**	**0.021**
Gemma	Base	0.000	0.670	0.007
	Tuned	**0.020**	**0.716**	**0.008**
Feedback				
OPT	Base	0.010	0.700	0.008
	Tuned	**0.036**	**0.741**	**0.010**
DGPT	Base	0.011	0.392	0.007
	Tuned	**0.015**	**0.689**	**0.101**
GEMMA	Base	0.023	0.700	0.012
	Tuned	**0.028**	**0.717**	**0.013**

Table 4.13: Automatic evaluation results ↑ indicates higher score is better.
DGPT: DialoGPT. **Bold** indicates the better value between base and fine-tuned models.
Underline indicates better metric in each feedback category.

4.3.2 Human Evaluation

Human evaluation serves in assessing the quality of generated feedback. This section will present the results of the human evaluation experiment conducted on the generated texts as described in Section 3.4.5.2 of Chapter 3 to gauge the appropriateness, correctness, and contextual relevance of generated outputs.

4.3.2.0.1 Participants Demography: The participant cohort for this analysis is characterized by a diverse age range and gender distribution. Among 49 participants, 57.1% identify as male while the remaining identify as female, showing a slightly higher representation of males. The majority of raters, 61.2%, fall within the 26-35 age range. The second largest age group is 36-45, comprising 18.4% of the participants. The smallest age group represented is 46-55, making up 4.1% of the participants. Overall, the pool skews toward younger adults, with nearly 80% of participants being under 36 years old. The data is represented graphically in Figures 4.16 and 4.17.

4.3.2.0.2 Inter Annotator Agreement (IAA): For each of the 7 surveys conducted, the IAA was quantified using Fleiss' Kappa, which measures the degree of consensus among multiple raters or annotators (Falotico and Quatto, 2015). Fleiss' Kappa values range from -1 to 1, with 1 indicating perfect agreement, and values above 0.6 generally considered substantial agreement. The average Kappa values from the surveys is calculated for each model. Table 4.14 presents

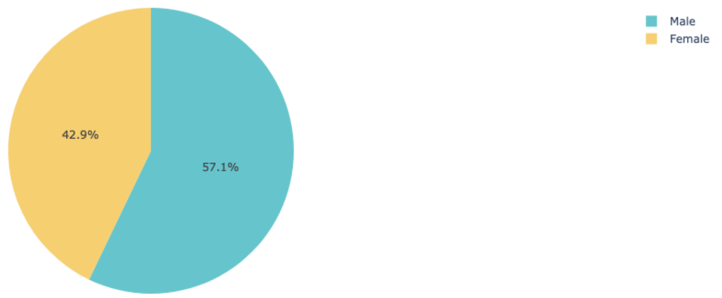

Figure 4.16: Distribution of participants by gender

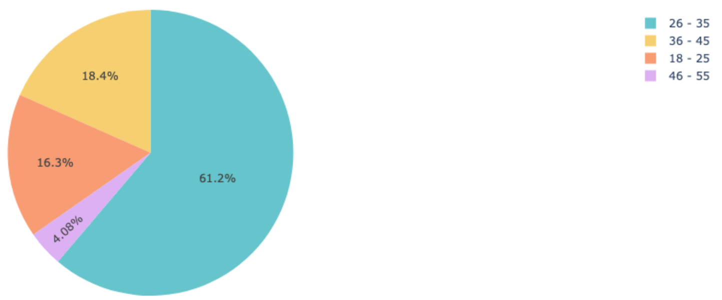

Figure 4.17: Distribution of participants by age

these results.

For Rephrasing Questions, the annotators exhibited slight to fair agreement. OPT and DGPT achieved slight agreement among annotators with scores of **0.067** and **0.113** respectively while Gemma achieved a fair agreement score of **0.258**. In the Feedback task, the IAA was slightly higher, with Fleiss' Kappa scores indicating slight to fair agreement. OPT and DGPT achieved **0.231** and **0.257** respectively. Gemma achieved the highest IAA of **0.347** in this category, suggesting fair consensus among the annotators.

	OPT	DGPT	Gemma
Rephrasing	0.067	0.113	**0.258**
Feedback	0.231	0.257	**0.347**
Overall	0.149	0.185	**0.303**

Table 4.14: Inter Annotator Agreement (IAA) scores. **Bold** indicates better agreement in each category.

4.3.2.1 Overall Performance of Models

Table 4.15 summarizes the overall performance of each model on the four evaluation criteria on the 5-point Likert scale. OPT achieved the highest rating across all categories with an average rating 2.9 out of 5.

For Fluency, OPT achieves the best rating of 3.1 with DialoGPT following closely with a rating score of 3.0, while Gemma received the lowest rating of 2.8 among the three models. In coherence, Gemma performed nearly as well as OPT, with only a 2% difference in their ratings while DialoGPT received the lowest rating of 2.4 for this criterion. Gemma again placed second behind OPT, achieving a rating of 2.1 in correctness. OPT stood out with a rating of 3.3 in appropriateness, while both DialoGPT and Gemma received a rating of 3.1.

	OPT	DialoGPT	Gemma
Fluency	**3.1**	3.0	2.8
Coherence	**3.0**	2.4	2.9
Correctness	**2.3**	1.9	2.1
Appropriateness	**3.3**	3.1	3.1
Overall	2.9	2.6	2.7

Table 4.15: Overall average rating of models for each question out 5.
DGPT: DialoGPT. **Bold** indicates the better value in each criteria.

4.3.2.2 Performance by Feedback Category

This section presents the performance of each model across the two categories.

4.3.2.2.1 Rephrasing Questions: In terms of fluency, OPT and DialoGPT performed equally well with a rating of 3.2, while Gemma lagged slightly behind with 2.9. For coherence, OPT and Gemma shared the highest rating of 3.0, outperforming DialoGPT's 2.6. In correctness, all three models struggled, with OPT receiving the highest score of 2.0 and DialoGPT and Gemma tied at 1.9. For appropriateness, OPT led with 3.3, followed closely by DialoGPT at 3.2 and Gemma at 3.0.

4.3.2.2.2 Feedback on Question Answer: For Fluency, OPT has the best average rating of 2.9, while DialoGPT and Gemma trailed with 2.8 and 2.6 respectively. In terms of coherence, OPT again led with 2.9, followed by Gemma at 2.7, and DialoGPT lagged behind with 2.0. Correctness was a challenge, with OPT performing best at 2.6, Gemma scoring 2.3, and DialoGPT lowest at 1.8. Finally, for appropriateness, OPT ranked highest with 3.2, Gemma second at 3.0, and DialoGPT third with 2.9. Similar to the trend seen in Rephrasing, OPT performs the best overall.

Overall, OPT and Gemma outperformed DialoGPT across both categories. However, all models struggled with correctness, especially in the rephrasing category. These resulted are presented in Table 4.16.

	FLN	CHR	COR	APR
Rephrasing				
OPT	3.2	3.0	2.0	3.3
DialoGPT	3.2	2.6	1.9	3.2
Gemma	2.9	3.0	1.9	3.0
Feedback				
OPT	2.9	2.9	2.6	3.2
DialoGPT	2.8	2.0	1.8	2.9
Gemma	2.6	2.7	2.3	3.0

Table 4.16: Average rating of models for each question out 5 by feedback category. FLN: Fluency. CHR: Coherence. COR: Correctness. APR: Appropriateness

4.3.2.3 Statistical Analysis

Table 4.17 presents the results of the ANOVA tests conducted to analyze the impact of evaluation on the evaluated criteria for each category of feedback.

In rephrasing, the test reveals non-significant differences across the four criteria since the p-values are greater than 0.05. Therefore, we fail to reject the null hypothesis since there is not enough evidence to to support the alternative hypothesis.

In contrast, for Feedback category, the p-values for coherence and correctness are less than 0.05 indicating that the result is statistically significant. However, no significant differences were observed in fluency ($F = 1.093, p = 0.338$) and appropriateness ($F = 0.886, p = 0.414$).

	F-value	**p-value**
Rephrasing		
Fluency	1.238	0.293
Coherence	2.149	0.120
Correctness	0.314	0.731
Appropriateness	0.236	0.790
Feedback		
Fluency	1.093	0.338
Coherence	7.438	**0.001**
Correctness	7.028	**0.001**
Appropriateness	0.886	0.414

Table 4.17: ANOVA test results of models

Chapter 5

Analysis and Discussion

This chapter presents an analysis and discussion of the findings obtained in Chapters 4. The chapter provides an examination of the results, situating them within the broader context of the research questions and hypotheses outlined.

5.1 Feedback Detection

In addressing the first research question regarding the detection of feedback strategies within child-adult dialogues. Four feedback categories are identified with the assistance of Linguistics Experts[1]. These include - Modelling Question Answer, Exemplifying Question Answer, Feedback on Question Answer and Rephrasing Questions. Two methods, a rule-based and an automatic detection approach are utilized for the detection.

The accuracy of the rule-based system for feedback detection, after validation against annotations from Linguistics Experts was **46%**. Upon comparison, the automatic detection approach demonstrated a notable improvement in accuracy, achieving an overall accuracy of **70%**. This shows a **24%** increase compared to the rule-based approach. There was increased accuracy from the rule-based approach in all feedback categories.

The computed perplexity scores show notable alignment between regions of higher perplexity values and the instances annotated as requiring feedback from adults. This aligns with the initial hypothesis that segments identified as surprising or unexpected by the language model would correlate with incorrect utterances by children, thus motivating the need for feedback. Additionally, in logistic regression classifier, perplexity ranks high among the most important features used for prediction for all feedback categories. This shows that language model properties can serve as additional features for the detection feedback.

To answer the first research question, feedback strategies of question-answers sequences between children and their caregivers can be detected, using both a rule based algorithm and an automatic detection method that leverages properties of LLMs.

[1] Linguistics Department at the University of Aberdeen

5.2 Feedback Generation

To address the second research question on generating contextually appropriate feedback for children, the three selected models: OPT, DialoGPT, and GEMMA were evaluated. This section will present an analysis of the results of the automatic and human evaluations conducted.

5.2.1 Automatic Evaluation

The automatic evaluation metrics provided valuable insights into the relative performance of the three fine-tuned models in generating contextually appropriate feedback responses. A detailed examination of the results revealed subtle variations in the capabilities of each LLM, as reflected in their scores on the three metrics.

For the reference based metrics, BLEU and BERT scores, which assess the similarity between the generated responses and the target outputs, OPT emerged as the top performer. Its higher BLEU scores indicate a higher degree of n-gram overlap with the reference responses, suggesting that OPT's generated feedback more closely resembled the expected outputs in terms of word choice and phrasing. Similarly, OPT's higher BERT scores point to a stronger semantic alignment between its generated texts and the target feedback, indicating a better grasp of the underlying context and meaning. On the other hand, DialoGPT performed the best in the Mauve metric, which compares the distribution of the generated text to the distribution of human-written text. This suggests that DialoGPT's responses align well with the statistical patterns of the target output. Interestingly, Gemma's performance fell slightly behind the other two models across all the metrics evaluated.

Overall, the automatic evaluation results underscore the complexity of the feedback generation task and the nuanced trade-offs that exist between different evaluation metrics. These findings suggest that an ideal feedback generation system may benefit from combining the strengths of multiple language models or employing ensemble techniques to leverage their complementary capabilities.

5.2.2 Human Evaluation

For the human evaluation, OPT model exhibited the lowest overall IAA of **0.149** and Gemma the highest overall with **0.303**. These results highlight the inherent subjectivity in the annotation tasks and the need for careful interpretation of the generated outputs. The slight to fair IAA scores suggest reasonable reliability of the annotations, albeit with room for improvement.

Human evaluations largely mirrored the trends observed in automatic metrics, reinforcing the key findings. Similar to its performance with the automated metrics, OPT emerged as the leader, surpassing the other models with an average human rating of **2.9** out of 5 as highlighted in Table 4.15. It's important to consider that while this score might appear moderate, it reflects the inherent difficulty of feedback generation and the challenge of creating responses that align with human expectations.

The statistical analysis shows that for the Rephrasing task, there is non-significant difference across the evaluation criteria for the three models however, in the Feedback category, there is statistical significance in the correctness and coherence. This helps draw the conclusion that OPT is significantly better than DialoGPT and Gemma for these two criteria. However, in the Feedback category, the statistical analysis revealed significant differences in the correctness and coherence criteria both achieving a p-value of **0.001**. This shows OPT's performance was significantly better than DialoGPT and Gemma in terms of producing feedback that was both correct and coherent within the given context. This finding shows OPT's strength in understanding the dynamics of the child-adult dialogue and generating feedback that not only accurately addresses the child's utterance but also maintains a logical flow and coherence.

For both evaluation methods, OPT performed the best. However, it is important to note that the language used in child-adult dialogues is often incoherent, which may explain why the language models performed only averagely. It was hypothesized that Gemma would perform well given its strong language understanding capabilities, but the results did not align with expectations.

5.2.2.1 Example Generations

The sample generation in Figure A.1 of Appendix A is discussed below:

In the context provided between a caregiver and child, it can be observed that the child makes an erroneous utterance in line 6 by responding with "horsie doing?" instead of a complete sentence. The caregiver promptly addresses this error in the next utterance by rephrasing the question as "what is the horsie doing?", providing an opportunity for the child to learn the correct form of the question. Each model generated response is analysed:

OPT: what's the horsie doing, The response effectively rephrases the caregiver's question, maintaining the same structure and meaning. It add a comma at the end.

DialoGPT: what is he doing? DialoGPT's response also rephrases the child's question. However, by using the pronoun "he" instead of "the horsie," the response may introduce some ambiguity or confusion for the child, as it is unclear whether the model is referring to the same subject (the horsie) or something else.

Gemma: horsie doing. what's Gemma's response appears to be less coherent and incomplete. While it repeats the child's erroneous utterance "horsie doing.", it does not provide a clear rephrasing of the child's question. The addition of "what's" at the end further contributes to the confusion, as it is unclear what the model is trying to convey.

Based on this sample, OPT's response appears to be the most effective and appropriate feedback, as it directly addresses the child's error by rephrasing the caregiver's question in a clear and concise manner. DialoGPT's response is also reasonable, although the use of the pronoun "he" could potentially introduce ambiguity. Gemma's response, however, seems to lack coherence and fails to provide clear feedback to the child.

In Appendix A, Tables A.1 and A.2 showcase the top-rated generations produced by each model across the two feedback categories while Tables A.3 and A.4 present the lowest rated generations produced by each model across the two feedback categories.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This project aimed to identify and detect feedback strategies present in child-adult dialogues, specifically in the context of question answer sequences and generate contextually appropriate feedback for children in such interactions.

To achieve this, both a rule-based algorithm and automatic approaches were employed for the detection task. The rule-based algorithm achieved reasonable accuracy, while the automatic detection mechanism, which utilized properties from LLMs, achieved an increased overall accuracy. The feature importance analysis on the automatic detection approach showed that perplexity of LLMs was an important feature for prediction. The results indicate that language model perplexity can serve as a useful signal for feedback detection systems.

The feedback generation experiment showed the potential of LLMs in generating contextually appropriate feedback for children. Three state-of-the-art language models - OPT, DialoGPT and Gemma were fine-tuned on a large corpus of child-adult dialogues. The three models were evaluated using both automatic metrics and human raters and it was found that OPT performed the best for both evaluation methods out of the three models.

By detecting different forms of question-answer feedback and testing the capabilities of LLMs in generating contextually appropriate feedback for children, this body of work contributes to the field of first language acquisition and natural language processing. The insights gained from this project can inform future research on language development and guide the design of educational interventions and tools that support children's linguistic growth.

6.2 Future Work

The findings from this project offer a promising foundation for future research. However, it is essential to acknowledge the limitations that this study encountered, which present opportunities for improvement to be explored by researcher in years to come. These areas are highlighted below:

- **Expanding Annotated Data:** The approach proposed by Nikolaus et al. (2024) can be utilized as a potential solution for automatically annotating the extracted feedback instances. Their methodology, which leverages machine learning techniques to assess grammaticality in child-caregiver conversations, holds promise for automating the annotation process for feedback sequences. This automated annotation approach not only reduces the time and effort required for manual annotation but also ensures consistency and scalability, enabling researchers to analyze larger volumes of data more efficiently.

- **Implementing an Better Classifier:** By using a larger annotated dataset, a more robust classifier can be developed in other to extend the automatic detection mechanism to the entire dataset, beyond the manually annotated data.

- **Better Fine-tuning Approach:** The instruction tuning Ouyang et al. (2022) method can be used to fine-tune LLMs for feedback detection. Instruction tuning, also known as fine-tuning with human feedback, is a technique used to align language models with user intent by training them to follow specific instructions provided by humans (Ouyang et al., 2022). With access to a large dataset containing labelled feedback instances, language models can undergo a targeted fine-tuning process by specifically instructing them to generate feedback when a child makes an erroneous utterance. This method holds the potential to enhance the models' understanding of the feedback generation task, enabling them to produce not just contextually relevant but tailored feedback responses.

Appendix A

A.1 Directories Structure

The directory contains the code base for the project. The directory is organized into the following main components:

```
childes-thesis
├── data
│   ├── extract-data
│   ├── prepare-train-data.ipynb
│   └── prepare-test-data.ipynb
├── feedback-detection
│   ├── extract-utterance-pairs.ipynb
│   └── logistic-regression.ipynb
├── train-generate
│   ├── train-opt-dgpt.py
│   ├── train-generate-gemma.ipynb
│   ├── opt-dgpt-generate.ipynb
│   └── opt-dgpt-base-model.ipynb
└── analysis
    ├── automatic-evaluation.ipynb
    ├── ethicsforms
    ├── human-evaluation.ipynb
    ├── human-evaluation-1.ipynb
    └── survey.pdf
```

A.1.0.0.1 Structure description:

- **childes-thesis:** The top-level directory of the project.
- **data:** parent folder containing dataset.
- **data/extract-data:** contains script for extracting data from xml files into csv files.
- **data/exploratory-analysis.ipynb:** notebook containing the code used for statistical analysis, data visualizations, and general exploration of the CHILDES data.

- **data/prepare-train-data.ipynb:** notebook containing the code used for preparing the raw data for training.

- **data/prepare-test-data.ipynb:** notebook containing the code used for preparing the raw data for testing.

- **feedback-detection/extract-utterance-pairs.ipynb:** notebook containing the rule base algorithm used for extracting feedback types.

- **feedback detection/logistic_regression.ipynb:** notebook containing the rule base algorithm used for extracting feedback types.

- **train-generate:** parent folder containing script and notebooks used for training and generation with both base and fine-tuned models.

- **analysis:** parent folder containing python notebooks used for automatic and human evaluation analysis.

- **analysis/survey.pdf:** sample survey form used for human evaluation.

- **analysis/ethicsforms:** survey participation consent form and related ethics application forms.

A.1.1 Required Libraries

- Hugginface Transformers (Wolf et al., 2019)

- Numpy (Harris et al., 2020)

- Pandas (Wes McKinney, 2010)

- Plotly (Plotly Technologies Inc., 2015)

- Matplotlib (Hunter, 2007)

- Seaborn (Waskom, 2021)

A.2 Human Evaluation

A.2.1 Sample Survey Questions

This is an example of a dialogue between a child of age **1 year 8 months** and their caregiver. At **line 6**, the child makes an error, and at **line 7** the caregiver corrects it. **Lines 8, 9, and 10** are model generated texts, attempting to correct the child's error. Please rate the following statements for each of the model generated responses.

1: CAREGIVER: that one's a kind of a triangle hat isn't it?
2: CAREGIVER: now what's happening?
3: CHILD: horsie.
4: CAREGIVER: horsie
5: CAREGIVER: what's the horsie doing?
6: CHILD: horsie doing?
7: CAREGIVER: what is the horsie doing?
8: GEN 1: what's the horsie doing,
9: GEN 2: what is he doing?
10: GEN 3: horsie doing.

Figure A.1: Example generation for human evaluation for Rephrasing Question

This is an example of a dialogue between a child of age **2 year 8 months** and their caregiver. At **line 6**, the child provides a response to a question he/she asked, at **line 8** the caregiver provides feedback confirming the child's statement.
Lines 8, **9**, and **10** are model generated texts attempting to provide feedback confirming the child's utterance. Please rate the following statements for each of the model generated responses.

1: CAREGIVER: hi friend hi friend hi friend.
2: CHILD: oh oh friend are you okay?
3: CHILD: and what's this?
4: CAREGIVER: you know what this is, right?
5: CAREGIVER: what?
6: CHILD: an ambulance.
7: CAREGIVER: an ambulance, right.
8: GEN 1: it's an ambulance, it's an ambulance
9: GEN 2: right. right, which does this kind of sound
10: GEN 3: it's an ambulance . what

Figure A.2: Example generation for human evaluation for Feedback on Question Answer

A.2.2 Top Rated Generations

Model	Feedback Response	Prompt	Model Response	Average Rating
OPT	where is Papa ?	A: oh a cheese sandwich ? A: would you eat it if I fixed it for you ? C: yeah A: I'll fix it for you but you must eat it . C: okay C: where Papa ?	Papa? well Papa's at work	4.07
DialoGPT	where's teddy gone ?	C: dodie . A: where's teddy ? A: where's teddy ? A: go get him . A: go get teddy . C: where teddy gone ?	where's teddy gone ?	4.86
Gemma	what is Fraser doing ?	A: who is that ? C: Gordon . A: what's he doing ? C: sitting on the couch . A: sure he is . C: what i Fraser doing ?	what's he doing ?	4.36

Table A.1: Highest rated generation samples for rephrasing question

Model	Feedback Response	Prompt	Model Response	Average Rating
OPT	that's it.	A: not instead of . C: look . C: yeah A: oh dear . A: bad cough . C: is that the wrong bit ? C: that's the right way .	yes that's the right way that's right.	4.46
DialoGPT	that's it.	A: not instead of . C: look . A: oh dear . A: bad cough . C: is that the wrong bit ? C: that's the right way .	-um. I'm sorry I didn't know	3.07
Gemma	another rivet.	A: close it. ? C: what's this ? A: that's a rivet . C: rivet . C: what's this ? C: a rivet .	that's a rivet. Rivet?	4.57

Table A.2: Highest rated generation samples for feedback on question answer

A.2.3 Lowest Rated Generations

Model	Feedback Response	Prompt	Model Response	Average Rating
OPT	where are the christmas cookies?	C: Fraser need something else ? A: what ? A: what Eve ? A: what dear ? C: Fraser want something else. C: where Christmas cookies ?	this week Fraser is gonna come and	1.7
DialoGPT	what are you doing?	A: oh look Carl . A: I didn't see that before . C: snake . A: it's a snake . C: Daddy . C: what you doing ?	you want, like that. It	1.5
Gemma	where are the christmas cookies?	C: where Christmas cookies ? A: what ? A: what Eve ? A: what dear ? C: Fraser want something else. C: where Christmas cookies ?	where Christmas cookies?	1.7

Table A.3: Lowest rated generations for rephrasing question

Model	Feedback Response	Prompt	Model Response	Average Rating
OPT	it does look like a gate , you're right .	A: crashing walls down . A: or crashing sand+castles down . A: anything you can crash . C: what's that , that look like ? A: I don't know , looks like a bunch of C: what's that ? C: that looks like a gate .	that's a house and a door. let's	1.93
DialoGPT	no this is my spoon .	A: maybe Fraser will get your crayons for you . C: no . A: okay Mom'll get your crayons . A: there they are . C: that little one . C: that ? C: that Fraser spoon .	-um -um -um what's	1.71
Gemma	that's it	A: not instead of . C: look . A: oh dear . A: bad cough . C: is that the wrong bit ? C: that's the right way .	we wadda put the	1.32

Table A.4: Lowest rated generation samples for feedback on question answer

Bibliography

Artstein, R. (2017). Inter-annotator agreement. *Handbook of linguistic annotation*, pages 297–313.

Belz, A. and Reiter, E. (2006). Comparing automatic and human evaluation of nlg systems. In *11th conference of the european chapter of the association for computational linguistics*, pages 313–320.

Bengio, Y., Ducharme, R., and Vincent, P. (2000). A neural probabilistic language model. *Advances in neural information processing systems*, 13.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT press.

Chouinard, M. M. and Clark, E. V. (2003). Adult reformulations of child errors as negative evidence. *Journal of child language*, 30(3):637–669.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Clark, E. V. (2020). Conversational repair and the acquisition of language. *Discourse Processes*, 57(5-6):441–459.

Clark, E. V. and Casillas, M. (2015). First language acquisition. In *The Routledge handbook of linguistics*, pages 311–328. Routledge.

Demetras, M. J., Post, K. N., and Snow, C. E. (1986). Feedback to first language learners: the role of repetitions and clarification questions. *Journal of Child Language*, 13(2):275–292.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dziri, N., Kamalloo, E., Mathewson, K. W., and Zaiane, O. (2019). Evaluating coherence in dialogue systems using entailment. *arXiv preprint arXiv:1904.03371*.

Falotico, R. and Quatto, P. (2015). Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49:463–470.

Ferjan Ramírez, N., Lytle, S. R., and Kuhl, P. K. (2020). Parent coaching increases conversational turns and advances infant language development. *Proceedings of the National Academy of Sciences*, 117(7):3484–3491.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., and Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of child language*, 16(3):477–501.

Ferreira, A. and Atkinson, J. (2008). Designing a feedback component of an intelligent tutoring system for foreign language. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 277–290. Springer.

Gemma, T., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., et al. (2024). Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Han, J., Kamber, M., and Pei, J. (2012). *Data mining: concepts and techniques*. The Morgan Kaufmann series in data management systems. Elsevier, 3 edition.

Hanawa, K., Nagata, R., and Inui, K. (2021). Exploring methods for generating feedback comments for writing learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9719–9730.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

Hiller, S. and Fernández, R. (2016). A data-driven investigation of corrective feedback on subject omission errors in first language acquisition. In Riezler, S. and Goldberg, Y., editors, *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 105–114, Berlin, Germany. Association for Computational Linguistics.

Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review*, 26(1):55–88.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2019). The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Jurafsky, D. and Martin, J. H. (2009). *Speech and language processing*. Prentice Hall, Pearson Education International, London [u.a.], 2. ed., [pearson international edition] edition.

Kantrowitz, M. (1994). Names corpus. https://www.kaggle.com/datasets/nltkdata/names.

Kochmar, E., Vu, D. D., Belfer, R., Gupta, V., Serban, I. V., and Pineau, J. (2020). Automated personalized feedback improves learning gains in an intelligent tutoring system. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21*, pages 140–146. Springer.

Laurence, S. and Margolis, E. (2001). The poverty of the stimulus argument. *The British Journal for the Philosophy of Science*, 52(2):217–276.

Lester, N. A., Moran, S., Küntay, A. C., Allen, S. E., Pfeiler, B., and Stoll, S. (2022). Detecting structured repetition in child-surrounding speech: Evidence from maximally diverse languages. *Cognition*, 221:104986.

Li, J., Galley, M., Brockett, C., Spithourakis, G. P., Gao, J., and Dolan, B. (2016). A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Liu, C.-W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., and Pineau, J. (2016). How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Loem, M., Kaneko, M., Takase, S., and Okazaki, N. (2023). Exploring effectiveness of gpt-3 in grammatical error correction: A study on performance and controllability in prompt-based methods. *arXiv preprint arXiv:2305.18156*.

MacWhinney, B. (2000). *The Childes Project: Tools for Analyzing Talk. Transcription format and programs*. CHILDES project. Lawrence Erlbaum.

Marti, P. and Giusti, L. (2010). A robot companion for inclusive games: A user-centred design perspective. In *2010 IEEE international conference on robotics and automation*, pages 4348–4353. IEEE.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Molnar, A., Jumelet, J., Giulianelli, M., and Sinclair, A. (2023). Attribution and alignment: Effects of local context repetition on utterance production and comprehension in dialogue. *arXiv preprint arXiv:2311.13061*.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Nagata, R. (2019). Toward a task of feedback comment generation for writing learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3206–3215.

Ng, A. and Jordan, M. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14.

Nikolaus, M., Agrawal, A., Kaklamanis, P., Warstadt, A., and Fourtassi, A. (2024). Automatic annotation of grammaticality in child-caregiver conversations. *arXiv preprint arXiv:2403.14208*.

Nikolaus, M. and Fourtassi, A. (2023). Communicative feedback in language acquisition. *New Ideas in Psychology*, 68:100985.

Novikova, J., Dušek, O., Curry, A. C., and Rieser, V. (2017). Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*.

O'Brien, C., O'Mara, M., Issartel, J., and McGinn, C. (2021). Exploring the design space of therapeutic robot companions for children. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 243–251.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Pillutla, K., Swayamdipta, S., Zellers, R., Thickstun, J., Welleck, S., Choi, Y., and Harchaoui, Z. (2021). Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

Plotly Technologies Inc. (2015). Collaborative data science.

Qian, Q., Huang, M., Zhao, H., Xu, J., and Zhu, X. (2017). Assigning personality/identity to a chatting machine for coherent conversation generation. *arXiv preprint arXiv:1706.02861*.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training. *OPENAI*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*.

Reiter, E. (2018). A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

Rivers, K. and Koedinger, K. R. (2013). Automatic generation of programming feedback: A data-driven approach. In *The First Workshop on AI-supported Education for Computer Science (AIEDCS 2013)*, volume 50, pages 50–59.

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child development*, 83(5):1762–1774.

Saint-Georges, C., Chetouani, M., Cassel, R., Apicella, F., Mahdhaoui, A., Muratori, F., Laznik, M.-C., and Cohen, D. (2013). Motherese in interaction: at the cross-road of emotion and cognition?(a systematic review). *PloS one*, 8(10):e78103.

Saxton, M. (2000). Negative evidence and negative feedback: Immediate effects on the grammaticality of child speech. *First Language*, 20(60):221–252.

Saxton, M., Backley, P., and Gallaway, C. (2005). Negative input for grammatical errors: Effects after a lag of 12 weeks. *Journal of child language*, 32(3):643–672.

Serban, I., Sordoni, A., Bengio, Y., Courville, A., and Pineau, J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Skinner, B. F. (1957). *Verbal behavior*. New York: Appleton-Century-Crofts.

Snow, C. E. (1972). Mothers' speech to children learning language. *Child development*, pages 549–565.

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. (2022). Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Tomasello, M. (2005). *Constructing a language: A usage-based theory of language acquisition*. Harvard university press.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Troussas, C., Papakostas, C., Krouska, A., Mylonas, P., and Sgouropoulou, C. (2023). Personalized feedback enhanced by natural language processing in intelligent tutoring systems. In *International Conference on Intelligent Tutoring Systems*, pages 667–677. Springer.

Van Der Lee, C., Gatt, A., Van Miltenburg, E., Wubben, S., and Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Voita, L. (2023). Word embeddings. Web page. Accessed: [date accessed].

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021.

Wes McKinney (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. (2022). Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019a). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. (2019b). Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.