Pragmatic Modelling in Language Learning: Caregiver Question-Answer Feedback in Child-Directed Dialogue

Maryam Bala[◁], **Johannes Heim**[▷], **Elspeth Edelstein**[▷], **Arabella Sinclair**[◇▷]
University of Southampton[◁], University of Aberdeen[▷], University College London[◇]
m.l.bala@soton.ac.uk, {johannes.heim|elspeth.edelstein}@abdn.ac.uk, arabella.sinclair@ucl.ac.uk

Abstract

In language development, children learn to form Question–Answer (QA) sequences through caregiver feedback that adapts dynamically to their evolving linguistic abilities. Using expert annotated child-caregiver interaction, we examine four feedback types that guide children's acquisition of adult-like QA behaviour: caregiver instructions through reformulating and affirming a child's output as well as caregiver demonstrations through exemplifying and modelling adult-like behaviour. Our analysis reveals that feedback incidence, frequency and complexity progress and adapt over the course of development, akin to a tailored curriculum for pragmatic development. We release our annotated dataset which offers a rich resource for studying pragmatic feedback and provides the first large-scale empirical evidence of adaptive, tailored caregiver feedback on QA behaviour. **Keywords:** Dialogue, Language Acquisition, Pragmatics, QA Feedback, Child development, Repetition

1. Introduction

Caregivers provide feedback to children in dialogue using strategies such as repetition, corrective input, and exposing them to adult conversation (Snow, 1977; Bohannon and Stanowicz, 1988). As children's linguistic competence grows, the nature and frequency of the feedback they receive should adapt accordingly (Snow, 1977), supporting a gradual progression from simple to more sophisticated language use (Wood et al., 1976). Prior research has primarily examined how caregiver feedback supports *syntactic* development (Marcus, 1993; Hiller and Fernández, 2016), leaving open questions about how context-specific modelling of adult behaviour contributes to the acquisition of *pragmatic* competence. Addressing this question is urgently needed because of growing evidence that language is learned in and through child-caregiver interaction (Clark, 2020). A promising phenomenon for studying pragmatic development is the acquisition of question-answer (QA) behaviour because adult-like QA behaviour requires an understanding of turn-taking, adjacency pairs, the relational overlap between questions and answers, and default mappings of clause types and speech acts. These properties have made question and their answers the central topic for understanding conversational dialogue, both in its development of shared beliefs (Stalnaker, 1978) and the relation between individual conversational moves (Roberts, 2012).

We therefore investigate caregiver feedback on early QA exchanges as a locus for understanding how pragmatic competence emerges via interaction. While previous investigations of caregiver feedback have targeted caregiver interventions and its relation to vocabulary growth (Ramírez et al., 2020) or morphosyntactic development (Hiller and Fernandez, 2016; Nikolaus et al., 2022), we propose that QA feedback represents an important window into the role of caregiver intervention on early pragmatic development. Our work expands the scope of prior work to include the distinction between functionally distinct feedback types, predicting that their onset, incidence and complexity will dynamically adapt to the learner's competence, thereby forming something akin to an unconscious curriculum of responsive input. Learning via exposure to a progression of increasingly complex examples has received considerable attention in the training of large language models under the notion of curriculum learning (Bengio et al., 2009), as has learning through feedback (Bai et al., 2022). If the goal is to develop more human-like models of language, then developmentally plausible reward signals offer a promising direction (Stöpler et al., 2025), enabling language models to learn through communicative success. Yet, an open question remains: Can systems that have been demonstrated to be successful for morphosyntactic and semantic output, also learn from developmentally plausible pragmatic feedback? While this question presents several challenges—particularly in identifying the different functions of pragmatic feedback as meaningful learning signal—our work takes an initial step toward addressing it by examining the nature and properties of pragmatic feedback in child language acquisition.

We thus annotate[1] and analyse a large collection of four functionally distinct QA feedback types (Heim et al., 2025) in early child-directed dialogue corpora on CHILDES (MacWhinney, 2000), showing their onset, incidence and complexity vary systematically with development in distinctive ways.

[1]Corpus and analyses at: https://github.com/the-context-lab/childQAfeedback

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11461–11478
11–16 May 2026.

We evaluate the ability of current large language models to recognize these pragmatic distinctions, finding that the LLM-as-a-judge approach, for the models we investigate, performs poorly on the nuanced pragmatic categories central to our study. Finally, while we create a comparatively large, rich resource of hand-annotated pragmatic feedback examples, we demonstrate that these can be useful to train automatic annotators that scale our data and enable us to analyse incidence and onset across CHILDES. We contribute a large-scale, expert annotated dataset of caregiver feedback QA sequences, from which we provide the first large-scale evidence that pragmatic caregiver feedback shows a curriculum-like, functionally diversified progression over early language development, suggesting a naturally structured sequence of how caregivers engage with their children's early pragmatic output.

2. Background

2.1. Child directed feedback

There is growing consensus that language learning happens in and through social interaction between caregivers and their children (Clark and Wong, 2002; Hoff, 2006; Rowe, 2012). Even approaches firmly believing in innate, language-specific capacities now acknowledge that caregiver input and its frequency play an important role in accounting for the stability and pace of language development (Yang et al., 2017). Input in the form of caregiver feedback is provided through various forms, such as corrections of non-adult child utterances (Demetras et al., 1986), repetitions (Bohannon and Stanowicz, 1988), reformulations (Chouinard and Clark, 2003), as well as corrective feedback (Hiller and Fernández, 2016) and communicative feedback (Nikolaus and Fourtassi, 2023), which constitute a multi-layered engagement from the caregiver with the child's learning process. Intensionally exposing children to complex constructions or recasting their output with targeted constructions, for instance, provide children with learning opportunities to adopt adult-like linguistic forms, aiding in syntactic and lexical development (Nelson, 1977). Likewise, a combination of recasts and corrections allow children to recognize and refine their errors (Farrar, 2020). These caregiver interventions help children learn the structural aspects of language as evidenced for both morphological (Clark, 2018) and syntactic (Hiller and Fernandez, 2016) development. However, we are only beginning to understand the role of these interventions for the development of pragmatic language skills, including turn-taking, conversational repair, and understanding social cues in communication (Dunn and Shatz, 1989; Clark, 2020; Nikolaus and Fourtassi, 2023). It is yet to be shown how these individual cues converge in early attempts of participating in dialogue.

We focus on QA behaviour because it captures a central ingredient of human interaction. Conversational dialogue is said to be structured by an ordered set of implicit or explicit questions that seek a shared understanding of what the (immediate) world is like (Roberts, 2012). The fabric of a conversation is therefore composed of conversational moves that establishes these questions and those that address them. The relations between these moves also explains how this fabric is held together: if the relation between question and answer is not obvious, interlocutors must negotiate the relevance of a contribution (Heim, 2025). Adult-like QA-behaviour therefore goes beyond understanding how to identify a correct question or answer; it requires an understanding of their relation, or congruence. What can aid the recognition of this functional congruence is an overlap in form. So-called information-questions (which include a wh-pronoun) are prime examples of lexical overlap between questions and their answers because they contain identical lexical strings (Reich, 2002). We exploit this form-based overlap to identify relevant instances of caregiver feedback on QA behaviour.

2.2. Mining dialogue interaction patterns

Large-scale examination of dialogue data from the CHILDES database (MacWhinney, 2000) has shed light on properties of child-caregiver interaction across language development. One characteristic of these dialogues is the between-speaker repetition between child and caregiver. Sinclair et al. (2021) observe that this repetition of both lexical and morphosyntactic constructions is highly local, as well as asymmetric; that is, the caregiver is far more likely to repeat the child than the child is to repeat the caregiver. This repetition can take the form of echoes, immediately adjacent repetition of the interlocutor's prior utterance (Takahashi, 1991), which this local finding captures.

Feedback will not always consist of direct repetition. Indeed, the characteristic of *corrective* feedback is in reformulation; of subject omission errors (Hiller and Fernández, 2016), development of speech acts (Nikolaus et al., 2021), dependency length minimization (Liu and Wulff, 2023), grammaticality (Nikolaus et al., 2024), and response contingency (Agrawal et al., 2024). A key challenge for investigating caregiver interventions of different forms lies in the identification of patterns across different pragmatic categories of interaction. These patterns cannot be captured through simple repetition-based or similarity metrics. Instead, much like intent recognition or speech/dialogue act classification, they require a more context-aware approach where

the preceding utterances (Raheja and Tetreault, 2019; Tanaka et al., 2019; Ahmadvand et al., 2019), speaker details, temporal information, and indeed punctuation (Malhotra et al., 2022; Żelasko et al., 2021) are often crucial for correctly determining the label of a given speech act. From a cognitive science standpoint, this context is essential, as it is equally important to humans when predicting dialogue act intents (Linders and Louwerse, 2023).

Influenced by more recent work demonstrating the effectiveness of pre-trained transformer architectures for exploiting context and structure in dialogue act (Żelasko et al., 2021) and sentence structure (Liu and Wulff, 2023), we explore effective approaches for interaction-sequence labelling—comparing a BERT-based (Devlin et al., 2018) prediction model to a simpler logistic regression model using linguistically-motivated semantic and information theoretic features—and develop a sufficiently effective automatic labelling technique to scale up our analyses.

3. Question-Answer Feedback

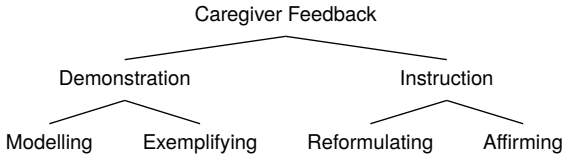

Figure 1: Taxonomy of caregiver feedback types.

We focus on caregiver feedback in response to early question–answer (QA) sequences in dialogue, including both responses to child attempts and adult-initiated modelling. All feedback types we examine relate to the child's developing understanding of the functional *congruence* of QA sequences, that is, how an answer appropriately satisfies the intent of a preceding question. Building on our work in (Heim et al., 2025), we focus on four functionally distinct types of caregiver feedback. These map on two broad pedagogical functions strategies: ***instructing*** the child on how to adjust their QA behaviour, and ***demonstrating*** adult-like QA behaviour.[2] Examples of each interaction type can be found in Table 1. [3]

Demonstration strategies Some caregiver feedback serves to supply the child with an example of an adult-like QA sequence. These can be either adult initiated; whereby the desired QA behaviour is modelled by the adult in response to an (extra)linguistic event, or it can be an adult extension of a child-initiated question attempt, whereby this question is extended to a full QA-sequence into an example. ***Modelling*** is thus defined as adult initiated behaviour where an adult demonstrates a full QA sequence; and ***Exemplifying*** as a child initiated behaviour where an adult repeats a child's question and provides an answer to it.

Instruction strategies Some QA feedback is more deliberate and directly responds to child's attempt at QA behaviour, often with some non-adult phrasing during the early stages. ***Reformulating***, is defined as feedback where an ill-formed question is reformulated by the adult. This feedback most closely matches other studies into corrective feedback and can serve as a counterpoint for morphosyntactic development. ***Affirming*** is defined as an adult confirmation of a child attempt at a QA sequence. This confirmation is often made through repeating parts of the answer or validating it, thereby providing indirect evidence for adult-like QA behaviour.

Adult Echoes We additionally examine adult-repetitions of child utterances, which also serve to ratify children's utterances (Clark and Bernicot, 2008). Echoing the ambient language plays a key role in first language acquisition in both children's output and caregiver feedback. We pay special attention to echoes, the direct repetition of a preceding child utterance by the adult. Echoes can, but not must coincide with exemplifying. This repetition can also be a recast with a perspective or pronoun change e.g. *C: what are you doing? A: what am I doing?*, or the partial repetition of the preceding utterance e.g *C: Where doggy? A: The doggy?*.

4. Data

We extract data from the English portion of the CHILDES Database (MacWhinney, 2000),[4] a resource of high-quality linguist-curated transcripts of naturalistic dialogues between adults and children of varying ages. We select only those transcripts involving children between the ages of 1 and 4 years old (12-48 months), that consist of only a single child interacting with no more than 5 adults. This resulted in 2,026 transcripts, of 46 children, with an

[2] Appendix A.2 provides formal definitions of the feedback categories and details of the annotation guidelines.

[3] Depending on the surrounding context, utterances can form part of multiple feedback types. e.g. Modelling could form part of Exemplifying: C: where Christmas cookies ? *A: where are the Christmas cookies ? A: they're all gone* ., Exemplifying can also consist of Reformulating e.g. *C: where Christmas cookies ? A: where are the Christmas cookies ?* A: they're all gone.

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

Modelling	Exemplifying	Reformulating	Affirming
A: What does the doggy say? A: Woof-woof.	C: What doing ? A: What's he doing ? A: He's putting that flower into some water .	C: What doing Nana? A: What's Nana doing?	C: What's this? C: Pyjamas. A: Pyjamas, right.
Adult gives an example of how to answer a question by asking a question, then answering it themselves.	*Child asks a question, and Adult repeats or reformulates the question and answers it*	*Child asks a question, then Adult rephrases it to give the corrected version of that question*	*Child asks a question and answers it themselves, Adult repeats the correct answer, to confirm it.*

Table 1: Caregiver Question-Answer feedback definitions and examples.

average of ∼1000±600 utterances per dialogue, of that ∼37% belonging to the child.

4.1. Corpus annotation & analysis

To facilitate the extraction of examples of each feedback category—as well as Adult Echos to compare their incidence and properties to the other implicit feedback types we consider—, we developed a rule-based approach with linguistic heuristics to distinguish QA feedback types. Key features included speaker sequence, punctuation patterns, and lexical overlap between utterances. This method yielded 82,409 Modelling, 1,296 Exemplifying, 5,540 Reformulating, and 591 Affirming candidates for human annotation.

A representative subset (∼400 per category) was manually annotated by the second and third authors, both academics & expert linguists. Our expert annotators were provided with an utterance sequence along with five preceding turns for context. This annotation process was done whilst defining and iteratively refining an annotation scheme to allow for less expert future annotation efforts. To assess guideline robustness, the first and last authors annotated a final sample—including previously agreed items plus an additional 200 per category—yielding 0.8 agreement with expert labels and 0.83 between annotators. Remaining uncertainties within the second batch of items were reviewed and resolved by an expert annotator. These results indicate that the guidelines are sufficiently robust, and yielded what we are satisfied are high-quality labels.

A total of 1401 Modelling, 595 Exemplifying, 1052 Reformulating and 591 Affirming feedback candidates were annotated, resulting in 250 Modelling, 238 Exemplifying, 284 Reformulating and 214 Affirming examples which make up our dataset. [5] An overview of the data can be found in Table 2.

5. Measures

We define the following measures, used to analyse the extracted feedback types in Section 7. These are also used in our rule-based filter, as well as to extract features for the classifiers in Section 6.

Vocabulary Overlap . To calculate the vocabulary overlap (VO) between two utterances, we exclude stop words[6] and punctuation, then determine VO as the ratio of shared words between the current utterance u_c and the previous utterance u_p.

$$\text{VO} = \frac{|u_c \cap u_p|}{|u_c|} \tag{1}$$

Perplexity We take perplexity, as estimated by GPT2 (Radford et al., 2019), as a useful proxy for how well-formed an utterance is: our hypothesis being that differences in perplexity in adjacent utterances can capture useful information for predicting feedback. We measure the perplexity $P(U)$ of an utterance U consisting of of words w—where N is the total number of words in the sequence and $P(w_i)$ is the probability of each word w_i as predicted by the model—as:

$$P(U) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log_2 P(w_i)\right)$$

We hypothesise that badly-formed 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. We compute the perplexity of an utterance given the 5 previous utterances preceding it.[7] We also compute the ***perplexity difference*** as the difference in perplexity in two adjacent utterances:

$$\Delta P(U_i, U_{i+1}) = P(U_i) - P(U_{i+1})$$

[5] To validate that the rule-based filtering does not exclude a substantial number of valid candidates, we manually annotated 100 discarded items and found only 0.25% were incorrectly excluded—sufficiently low to justify the efficiency gains.

[6] For this analysis, the list of stop words used is based on the compilation outlined by (Hiller and Fernández, 2016), which represent the 100 most commonly occurring words within the CHILDES dataset (MacWhinney, 2000).

[7] Details can be found in Appendix B.1.

	M		E		R		A	
Age	0-2	2-4	0-2	2-4	0-2	2-4	0-2	2-4
Pos. examples	99	151	83	155	145	139	30	184
Neg. examples	510	642	67	289	236	533	49	329
Utt. length								
Adult	7.54 ± 3.15	9.97 ± 4.59	8.43 ± 2.99	10.92 ± 5.17	4.35 ± 1.67	5.07 ± 2.37	4.60 ± 3.86	6.38 ± 4.73
Child	-	-	3.16 ± 1.44	4.58 ± 2.14	2.90 ± 1.38	4.32 ± 2.24	5.0 ± 2.13	7.97 ± 4.05
VO	0.13 ± 0.26	0.11 ± 0.22	0.50 ± 0.17	0.47 ± 0.22	0.46 ± 0.19	0.56 ± 0.19	0.39 ± 0.17	0.39 ± 0.21

Table 2: QA Feedback properties. Utterance length (Utt len) of child and adult utterances, and Vocabulary Overlap (VO). M: Modelling. E: Exemplifying. R: Reformulating. A: Affirming. In total our dataset comprises 3639 annotations (both Pos. and Neg.), of which 986 are examples of the feedback we analyse.

such that a high positive value of perplexity difference indicates the later of the two is the less surprising (e.g., where a child utters a poorly formed question with high perplexity, and the adult corrects with a well-formed alternative with lower perplexity, ΔP is positive).

Cosine Similarity We compute the sentence embeddings of all utterances in the feedback pair or triple using Sentence-BERT (SBERT) (Reimers and Gurevych, 2019). We then calculate pairwise cosine similarity of adjacent utterances to assess the semantic relationship between utterance pairs.

Lexical Sophistication As a measure of lexical sophistication, we calculate the average age of acquisition for each utterance by averaging the age-of-acquisition ratings of its constituent words, using norms for 30,000 English words (Kuperman et al., 2012). Where informal or badly formed words were used, and thus not present in the norm data, these were not counted.

POS proportion To compute part-of-speech (POS) proportions for each child utterance, we extract and count POS tags, including common contractions (e.g., can't, won't) as negation, which are not explicitly tagged in CHILDES (MacWhinney, 2000). Counts are then normalized by the total number of tags to yield relative POS proportions.

6. Scaled up Feedback Annotation

To enable a full-scale annotation of CHILDES (MacWhinney, 2000), we investigate automatic annotation approaches which use examples from our annotated corpus.

6.1. Automatic Annotation Approaches

Rather than training a multi-class classifier, which may rely most on the signal present in the unique speaker sequences for each feedback type, we opt to train individual binary classifiers to distinguish between positive and negative feedback examples with the same speaker sequence. While this is not the most general approach, it leads to the highest-quality classifiers, which is our aim.

Logistic regression Four distinct binary classifiers are developed, each corresponding to a feedback category. To ensure balanced class representation and mitigate potential biases, equal samples of correct and incorrect feedback instances are included for each category. As features, we calculate the perplexity, perplexity difference, cosine similarity, and vocabulary overlap between all adjacent utterance pairs in the feedback sequence, aiming to capture some of the key heuristics used during human annotation.

BERT and ModernBERT We make use of our data to fine-tune BERT (Devlin et al., 2018) and ModernBERT (Warner et al., 2024) to create binary classification models specific to each feedback category e.g. *feedback-type vs other*. We use the Adam optimizer with a learning rate of 2×10^{-5}, binary categorical cross-entropy as the loss function, and train for 10 epochs, $\sim$1h of compute per run. For input, we tokenize the concatenated utterances and speaker labels of each feedback instance, and the labels are C *child*, and A *adult*. We include balanced negative examples for each feedback type. We also experiment with including a prior context utterance.

Prompting Finally, we make use of our dataset to construct few-shot prompts from the labelled examples. We explore prompting using multiple performant Instruction-tuned LMs: Llama 3 8b-Instruct (Grattafiori et al., 2024), Gemma 2b-Instruct (Team et al., 2024), and Falcon 7b-Instruct (Almazrouei et al., 2023).

We experiment with the following settings:

$Prompt_{def}$ We provide the annotation instructions for a particular target feedback category. We specify to answer yes or no.

$Prompt_{def+pos}$ In addition to Prompt$_{def}$, we provide a hand-chosen *positive* example of the candidate feedback type.

		M				E				R				A				Av.
		Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc
BERT	Base	-	-	-	-	0.51	0.51	0.98	0.67	0.50	0.50	1.00	0.67	0.49	0.00	0.00	0.00	0.50
	Tuned	-	-	-	-	**0.85**	0.80	0.94	0.86	**0.79**	0.73	0.93	0.82	0.74	0.74	0.74	0.74	0.79
BERT + C	Base	0.48	0.00	0.00	0.00	0.50	0.50	1.00	0.67	0.55	0.80	0.14	0.24	0.50	0.50	0.21	0.30	0.51
	Tuned	**0.75**	0.82	0.64	0.72	0.78	0.74	0.88	0.80	0.78	0.79	0.77	0.78	0.70	0.68	0.74	0.71	0.79
MBERT	Base	-	-	-	-	0.45	0.41	0.23	0.29	0.55	0.53	1.00	0.69	0.47	0.40	0.14	0.21	0.49
	Tuned	-	-	-	-	0.83	0.78	0.94	0.85	0.79	0.80	0.77	0.79	**0.76**	0.69	0.98	0.81	0.79
MBERT + C	Base	0.52	0.54	0.30	0.39	0.52	0.67	0.08	0.15	0.31	0.32	0.34	0.33	0.57	0.54	0.91	0.68	0.48
	Tuned	0.56	0.69	0.22	0.33	0.67	0.90	0.96	0.74	0.73	0.96	0.47	0.64	0.72	0.68	0.81	0.74	0.73
Logistic Regression		0.57	0.58	0.57	0.54	0.72	0.72	0.72	0.72	0.67	0.68	0.67	0.67	0.70	0.60	0.51	0.45	0.67
Llama3 8b-Instruct	$Prompt_{def}$	0.57	0.60	0.39	0.47	0.56	0.55	0.65	0.60	0.69	0.62	1.00	0.76	0.64	0.61	0.79	0.69	0.62
	$Prompt_{def+pos}$	0.48	0.40	0.09	0.14	0.66	0.61	0.90	0.72	0.66	0.59	1.00	0.74	0.64	0.61	0.77	0.68	0.61
	$Prompt_{def+pos+neg}$	0.48	0.44	0.17	0.25	**0.71**	0.64	0.98	0.77	0.67	0.60	1.00	0.75	0.63	0.59	0.86	0.70	0.62

Table 3: Evaluation results across feedback types with average accuracy. M: Modelling. E: Exemplifying. R: Reformulating. A: Affirming. Acc: Accuracy. Prec: Precision. Rec: Recall. F1: F1-score. **Bold** and underline indicates the best overall model accuracy for each feedback category.

$Prompt_{def+pos+neg}$ In addition to Prompt$_{def+pos}$, we provide an additional hand-chosen *negative* example of the candidate feedback type.

6.2. Results

We find that our **Bert-based** classifiers are the most performant (See Table 3 for an overview of the results). Interestingly, we do not find that MBert consistently outperforms Bert as we expected; indeed, it only performs best in one case. We explore the effects of including an additional contextual utterance[8], and again find that this only improves results for certain categories. We select the most performant models for our automatic annotation *Modelling:* $Bert_{+C}$, *Exemplifying:* $Bert$, *Reformulating:* $Bert$ and *Affirming:* $MBert$.

We further examine the results of these models broken down by how complex it is to annotate (see Table 9 in Appendix G). We make use of the human-annotation labels to group items into *easy*, where annotators both agreed upon the label with no discussion, and *hard*, where initial annotations were not the same and labels were resolved via discussion with an additional annotator. We observe that the accuracies are consistently higher for these less-ambiguous cases for Modelling and Affirming, with very little difference for Exemplifying, and mixed behaviour across models for Reformulating. We expect that the outcome of the automatic annotation will allow us to understand the patterns present in the most straightforward feedback examples, but may not allow us to analyse more complex nuanced cases without further human-annotation effort. For our logistic regression model, we observe, in line with our expectations and annotation experience: for Reformulating and the initial utterance pair in Exemplifying, the most predictive features are cosine similarity and VO between utterances. A similar pattern emerges for Modelling and Affirming, though the effect is less pronounced. We note that our logistic regression model performs on-par with, if not better than most of our experiments with prompting larger instruction tuned models. We experimented with a variety of different prompts, following standard formatting guidelines, and found mixed results when including positive and negative examples across categories. The best performing LLM-as-a-judge model was Llama3 8b, which only performed marginally better than our logistic regression model for the prompt with no examples.[9]

7. Analysis: Changes in Feedback with Child Development

Questions and answers are central to conversation (Roberts, 2012), often receiving encouraging and positive feedback from caregivers at their children's early attempts. Using our gold and scaled up automatic annotations (Section 6), we examine how feedback categories vary with children's development. Since we hypothesise that caregiver intervention adapts to the child's linguistic ability, as seen in previous analyses of corrective feedback, we use *Reformulating* as a reference category against which to compare other feedback types.

7.1. Incidence of feedback types vary over time

We expect the proportion of feedback types that caregivers provide to children will vary as the child develops their linguistic competence (Heim et al., 2025). As shown in Figure 2, we observe a greater

[8]For the Modelling category we decided to always include a context utterance, since this was one of the annotation guidelines relied upon to resolve ambiguous cases in the final annotation stage.

[9]Results of feature importance for the logistic regression models as well as results of our other LLM-as-a-judge experiments can be found in Appendix G.

and earlier incidence[10] of Modelling and Reformulating, and lower incidence of Exemplifying and Affirming. In line with our predictions, paired t-tests with Holm-Bonferroni correction of p-values for multiple tests find significant differences across age-bins. We observe a greater proportion of Modelling in under-2-year-olds, reducing with age; a significant increase in Exemplifying as children reach 2 years of age; a significantly higher incidence of Reformulating in children between 20 and 30 months old, with the arrival of (complex) syntax; and a significant increase over development with Affirming feedback as child competence grows.

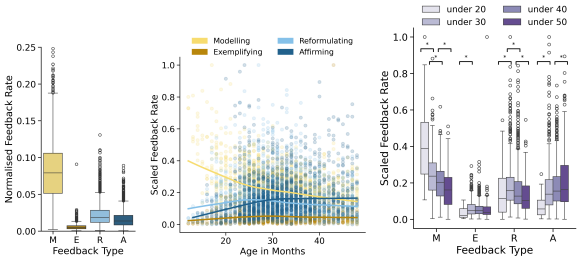

Figure 2: Feedback rate over child development. Min-max scaled rate of incidence for easier onset comparison. * shows significant differences in age bins.

7.2. Relationship between child complexity and feedback over development

To explain and provide context for the feedback rate results of Figure 2, we now investigate the relationship between the complexity of the feedback—or the complexity of the question attempt and its resulting feedback—and child development.

Utterance Length In line with child development, we observe a steady increase in child utterance length with age. In terms of the adult's complexity, for Exemplifying and Affirming, with increasing competence, the caregiver expands to include more lexical material with increasing age (e.g., Lustigman and Clark, 2019), indicating the caregiver is more nuanced in how they correct the child. That is, adults are more likely to reformulate or paraphrase child attempts rather than directly repeat back to them.

Figure 3 shows that mean length of utterance significantly increases for children over time (see Table 4, MLU), that is their attempts of asking and responding to questions increase in complexity with development, and that in their response to these attempts adults remain more static. For Modelling, however, we observe a significant increase in complexity of adult example question answer sequences, in-keeping with the observation of continued modelling in later stages, it is likely that these examples become more complex.

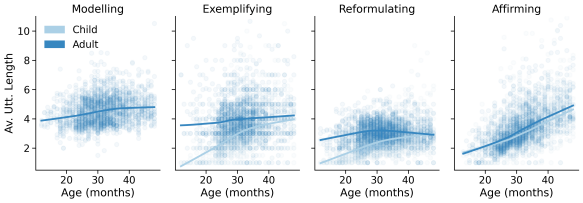

Figure 3: Utterance length by age across feedback type.

Lexical Sophistication As with utterance length, the complexity of vocabulary grows with child development, in line with expectations. As shown in Figure 4 the relationship between the vocabulary sophistication of child and adult affirms that caregiver feedback grows in complexity with output of the child, that is, the more sophisticated the child gets the more sophisticated the feedback is. Comparing this to the lexical sophistication observed in adult echoing (which also contain direct adult repetition of child question, see Section 7.3), we observe that in Exemplifying and Reformulating, caregivers match child level. We interpret this as caregivers adapting their feedback input to child ability (Lustigman and Clark, 2019).

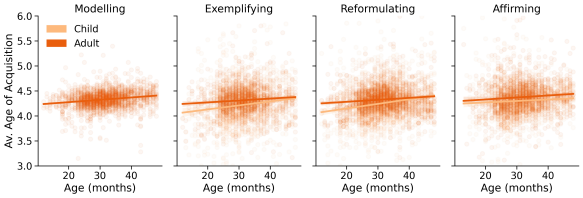

Figure 4: Lexical sophistication (average per-word Age of Acquisition) by age across feedback types.

Part of Speech Distribution We also analyse the incidence and proportion of parts of speech present in the child utterances based on the morphological tags provided in CHILDES, and how these vary across stages of child development. While Figure 2 contrasts trends of pragmatic versus reformulating feedback incidence, syntactic development may still constrain the child in expressing adult-like QA behaviour. POS proportion within the child question attempts can provide further context to understand the trends we observe, and the relationship with the complexity of the child attempt. It is only at later stages of child development that children will begin to use auxiliaries and negations, while determiners are a good indicator of syntactic and

[10] Normalised incidence rate: number of feedback instances in a dialogue normalised by feedback opportunities, which we define as dialogue length in utterances - (feedback length in utterances - 1).

pragmatic complexity, which help contextualise our observations on QA behaviour.

In keeping with expectations, in Figure 5 we observe significant increase in child use of Auxiliary and Verb usage, most so for Aux. in Exemplifying, and most so for Verb in Affirming. We also observe a steady increase for Determiner in Exemplifying and Affirming. In terms of negation this has a later increase in for Affirming. [11]

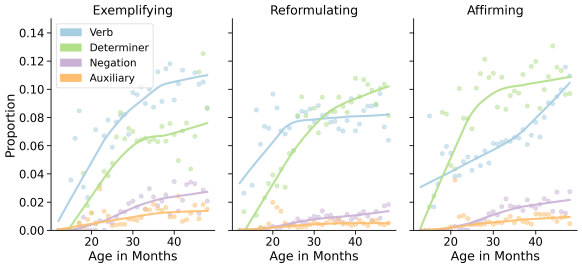

Figure 5: Proportion of child POS tags across development. Points are average value per age.

7.3. Repetition as feedback

A key aspect of all QA feedback is the extent and frequency of the repetition present across development. We examine both lexical and syntactic repetition, distinguishing between repetition between- vs. within-speaker. Between-speaker repetition is interpreted as reflective of feedback style, while within-speaker repetition may signal question–answer cohesion, shedding light on the emergence of QA congruence. Figure 6 shows that lexical repetition declines with development in both Modelling and child QA attempts, suggesting a shift toward more abstract question–answer relations.[12] These findings suggest that corrected question attempts grow in complexity (Figure 3), while corrections themselves become less precise in their repetition (Figure 6), reflecting competence-matching—caregiver adaptation to children's increasing linguistic sophistication (Lustigman and Clark, 2019).

Adults are more likely to reformulate or paraphrase child attempts rather than directly quote them (Lustigman and Clark, 2019). As a result, we observe a peak in between speaker VO—indicating a decrease in direct repetitions—with only minimal change overall, which aligns with expectations. The within-speaker repetition, which we take as the degree of simplicity between QA pairs: e.g., early

[11] Not shown is the proportion of Nouns and Adjectives, which exhibit a decrease in proportion as children learn a wider range of syntax, then a more gradual settling to adult-like proportions (see Table 4 for r coefficients).

[12] We also investigate Syntactic repetition, which shows a complementary trend, reflecting children's progression toward adult-like syntactic differentiation (Figure 10, Appendix H).

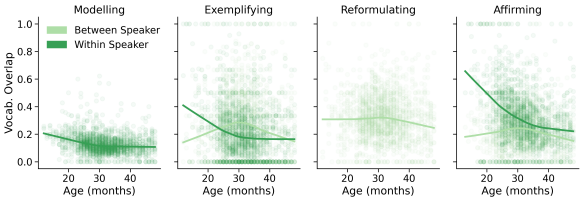

Figure 6: Vocabulary Overlap changes between and within speaker across feedback categories.

child QA attempts can look like Q:*dog?*, A:*it dog*, while later attempts may look more like Q:*where doggy going?*, A:*gone fido gone*. Thus, we see a sharp decline in Affirming, indicating the attempts receiving affirmation grow in complexity.

Adult Question Echos Echos of child questions by the caregiver can serve to signal what is considered common ground or given information (Schiefelin et al., 1979), to ratify the child's question (Clark and Bernicot, 2008), to signal agreement without taking up the floor in conversation (Yngve, 1970), to indicate information uptake, or to express confirmation (Tannen, 2007). What we expect, therefore, is heavy reliance on repetition for multiple uses, with a decrease in frequency and more specialisation with increasing vocabulary size (Kuhl and Meltzoff, 1996). We observe in Figure 7 that adults extend and expand on the child questions, both in length and lexical complexity, this shows a distinct pattern from the question repetition patterns present in Exemplifying and Reformulating, where the lexical sophistication of the child approaches the adults, and that the length increase of the adult repetition is more steady. This comparison provides further evidence that Adult feedback is adaptive to child ability.

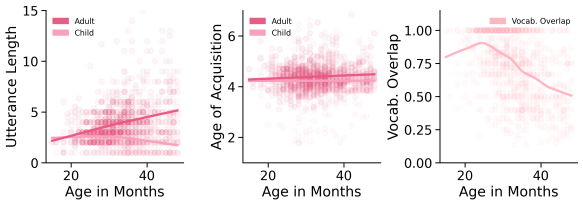

Figure 7: Adult echos of child questions.

8. Discussion & Conclusion

Our findings support the view that pragmatic QA feedback provides a highly adaptive learning environment for shaping children's understanding of how to recognize and respond to questions. Building on accounts that situate language learning in collaborative interaction (e.g. Clark, 2018), this study makes several key contributions: We provide the first large-scale evidence that pragmatic caregiver feedback occurs in distinct forms and shows

Property	M	E	R	A
MLU				
child	-	0.470	0.494	0.509
adult	0.231	0.090	0.036	0.582
AoA				
child	-	0.116	0.159	0.056
adult	0.191	0.089	0.100	0.072
POS Prop.				
Verb	-	0.073	0.011	0.094
Adj	-	-0.031	-0.026	-0.036
Noun	-	-0.169	-0.139	-0.153
Aux	-	0.056	0.004	0.021
Det	-	0.069	0.085	0.039
Neg	-	0.077	0.040	0.077
VO				
between	-	-0.031	-0.058	-0.104
within	-0.224	-0.181	-	-0.329
Struct. Dist				
between	-	0.182	0.255	0.213
within	0.176	0.146	-	0.514

Table 4: Pearson correlation coefficients (r). Insignificance ($p > 0.05$) in light gray. Mean length of utterance (MLU), lexical sophistication (AoA), POS proportion, vocabulary overlap (VO) and Structural Distance (levenshtein edit distance between- or within-speaker utterances) with respect to child age in months, by feedback and speaker.

a curriculum-like progression of child-adapted complexity over early language development, suggesting a naturally structured sequence of learning. Caregiver feedback not only adapts to the children's ability in type - with Modelling setting in before children ask multiword questions, and Affirming engaging with children's first QA sequences - but also in terms of the complexity of the feedback itself, which increases as a function of the growth in children's utterance complexity.

We present a comparatively large-scale, expert annotated dataset of caregiver feedback QA sequences (~ 1000 positive examples, from ~ 3500 total annotated samples). This resource captures fine grained pragmatic distinctions and can support future research into pragmatic feedback in humans and for evaluating language models for nuanced intent detection in dialogue. Finally, while task-specific classification models demonstrate moderate performance in classifying feedback types, our results highlight limitations in using language models to make judgements about such nuanced pragmatic distinctions. This resource can therefore serve as a useful future test for identifying pragmatic functions within dialogue interactions.

Our work provides new evidence of the importance of caregiver feedback through demonstration in the form of *Exemplifying* and *Modelling* feedback—acting as role models who offer examples of successful QA behaviour, a form of support rarely focussed upon in language development research. While we initially hypothesised that Modelling would decline once children reached a certain complexity threshold, it remains frequent even at later ages whilst becoming more complex: syntactic and lexical complexity increases, and the relationship between questions and answers becomes less explicit. We observe that later instances of Modelling gradually adopt functions familiar from adult dialogue where speakers use questions to confirm the topic of discussion, which are then self-answered (Bolden, 2009). Modelling thus not only aids child development but also plays an active role in shaping their understanding of discourse coherence. The fact that caregiver feedback aligns with child output complexity suggests a proactive—or at least subconsciously adaptive—modelling to the child's level, highlighting its relevance to language acquisition.

One additional outcome of this work is an increased interest and attention to *echoes* and pragmatic repetition in dialogue. Future work can investigate echoes and repetition in more depth to investigate differences between adult and child echo behaviour: Children often repeat back to their caregiver a new label for an object that they attend to, which marks an uptake of new word or information (Clark, 2010; Clark and Wong, 2002), or to endorse an interpretation by their caregiver of what the child tried to express earlier (Veneziano et al., 1990).

Building on Lustigman and Clark (2019) and (Nikolaus et al., 2022), taking in the distinctions between pragmatic variables contributing to question interpretations (Heim and Wiltschko, 2020), we develop and refine an annotation scheme and framework for extracting candidate instances. Identifying this feedback signal in language can be useful for understanding the nuance present in caregiver input to children, while also offering an alternative source of feedback when considering LM language learning; indeed, our findings provide motivation for a curriculum of pragmatic feedback, one potential direction for providing cognitively plausible signal for language learning that could inform the training regimes of more human-like LMs such as in (e.g. Stöpler et al., 2025; Zhu et al., 2022).

Acknowledgements

We would like to thank our anonymous reviewers for their thoughtful feedback, and have incorporated their suggestions in our work. JH and MB thank the audience at LAGB 2024 for their discussion and feedback for their feedback on our initial schema.

Limitations

Our work is limited to the English portion of CHILDES, thus our findings are only validated for the English language. In this work we focus our analyses on the full corpus of automatically annotated samples: while we are relatively satisfied with the precision of our classifiers, there will be some margin for error, which affects our analyses. A further limitation lies in the comparatively small number of manually annotated positive examples, especially for certain sub-categories (0-2 years, Affirming, see Table 2. A fundamental assumption of our approach is that the feedback sequences we analyse conform to a specific sequence of speaker utterances: this may result in excluding some sequences of slightly different composition which could also display the same properties.

Since the nature of this work was incremental and involved various stages of annotation and experimentation, our initial experiments comparing classifier performance were reported for a slightly smaller test set, and we note a degree of variability in our results with our new expanded set. We release our train-test splits that we use for replicability and advise in future work that results be reported for a cross validated average.

We also recognise that our results for LLM as a judge may be low due to the size of the models, in future work it would be interesting to investigate whether better results are achieved with larger models, or different configurations of judges.

Bibliographical References

Abhishek Agrawal, Mitja Nikolaus, Benoit Favre, and Abdellah Fourtassi. 2024. Automatic coding of contingency in child-caregiver conversations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1856–1870.

Ali Ahmadvand, Jason Ingyu Choi, and Eugene Agichtein. 2019. Contextual dialogue act classification for open-domain conversational agents. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 1273–1276.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, et al. 2023. Falcon-40b: an open large language model with state-of-the-art performance.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

John N Bohannon and Laura B Stanowicz. 1988. The issue of negative evidence: Adult responses to children's language errors. *Developmental psychology*, 24(5):684.

Galina B Bolden. 2009. Beyond answering: Repeat-prefaced responses in conversation. *Communication Monographs*, 76(2):121–143.

Peter Brodsky and Heidi Waterfall. 2007. Characterizing motherese: On the computational structure of child-directed language. In *Proceedings of the annual meeting of the cognitive science society*, volume 29.

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

E. Clark. 2018. Conversation and language acquisition. *Language Learning and Development*, 14(3):170–185.

Eve V Clark. 2010. Adult offer, word-class, and child uptake in early lexical acquisition. *First language*, 30(3-4):250–269.

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

Eve V Clark and Josie Bernicot. 2008. Repetition as ratification: How parents and children place information in common ground. *Journal of child language*, 35(2):349–371.

Eve V Clark and Andrew D-W Wong. 2002. Pragmatic directions about language use: Offers of words and relations. *Language in Society*, 31(2):181–212.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Judy Dunn and Marilyn Shatz. 1989. Becoming a conversationalist despite (or because of) having an older sibling. *Child development*, pages 399–410.

Michael Jeffrey Farrar. 2020. Negative evidence and grammatical morpheme acquisition. In *Language in Use*, pages 295–312. Routledge.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

J. Heim and M. Wiltschko. 2020. Deconstructing questions: Reanalyzing a heterogeneous class of speech acts via commitment and engagement. *Scandinavian Studies in Language*, 11(1):56–82.

Johannes Heim, Maryam Bala, and Arabella Sinclair. 2025. Modelling and feedback in the context of early questions. *PsyArXiv preprint*.

Johannes M Heim. 2025. Negotiating truth and relevance: A new typology of english rising declaratives. *Journal of Pragmatics*, 249:23–43.

S. Hiller and R. Fernandez. 2016. A data-driven investigation of corrective feedback on subject omission errors in first language acquisition. In *Proceedings of the 20th SIGNLL*, pages 105–114,, Berlin, Germany.

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

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

Patricia K Kuhl and Andrew N Meltzoff. 1996. Infant vocalizations in response to speech: Vocal imitation and developmental change. *The journal of the Acoustical Society of America*, 100(4):2425–2438.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, 44:978–990.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Guido M Linders and Max M Louwerse. 2023. Surface and contextual linguistic cues in dialog act classification: A cognitive science view. *Cognitive Science*, 47(10):e13367.

Zoey Liu and Stefanie Wulff. 2023. The development of dependency length minimization in early child language: A case study of the dative alternation. In *Proceedings of the seventh international conference on dependency linguistics (depling, gurt/syntaxfest 2023)*, pages 1–8.

L. Lustigman and E. Clark. 2019. Exposure and feedback in language acquisition: adult construals of children's early verb-form use in hebrew. *Journal of Child Language*, 46(2):241–264.

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

Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 735–745.

Gary F Marcus. 1993. Negative evidence in language acquisition. *Cognition*, 46(1):53–85.

Keith E Nelson. 1977. Facilitating children's syntax acquisition. *Developmental psychology*, 13(2):101.

M. Nikolaus, L. Prévot, and A. Fourtassi. 2022. Communicative feedback as a mechanism supporting the production of intelligible speech in early childhood. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44(44)).

Mitja Nikolaus, Abhishek Agrawal, Petros Kaklamanis, Alex Warstadt, and Abdellah Fourtassi. 2024. Automatic annotation of grammaticality in child-caregiver conversations. *arXiv preprint arXiv:2403.14208*.

Mitja Nikolaus and Abdellah Fourtassi. 2023. Communicative feedback in language acquisition. *New Ideas in Psychology*, 68:100985.

Mitja Nikolaus, Juliette Maes, Jeremy Auguste, Laurent Prevot, and Abdellah Fourtassi. 2021. Large-scale study of speech acts' development using automatic labelling. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Vipul Raheja and Joel Tetreault. 2019. Dialogue act classification with context-aware self-attention. *arXiv preprint arXiv:1904.02594*.

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

Ingo Reich. 2002. Question/answer congruence and the semantics of wh-phrases.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Craige Roberts. 2012. Information structure: Towards an integrated formal theory of pragmatics. *Semantics and pragmatics*, 5:6–1.

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

Bambi Schieffelin, E Ochs, and Martha Platt. 1979. Propositions across utterances and speakers. In *Developmental pragmatics*, pages 251–268. Academic Press.

Arabella J Sinclair, Raquel Fernández, et al. 2021. Construction coordination in first and second language acquisition. In *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue*.

Catherine E Snow. 1977. The development of conversation between mothers and babies. *Journal of child language*, 4(1):1–22.

Robert C. Stalnaker. 1978. Assertion. In Peter Cole, editor, *Syntax and Semantics 9: Pragmatics*, pages 315–332. Academic Press, New York.

Lennart Stöpler, Rufat Asadli, Mitja Nikolaus, Ryan Cotterell, and Alex Warstadt. 2025. Towards developmentally plausible rewards: Communicative success as a learning signal for interactive language models. *arXiv preprint arXiv:2505.05970*.

Mari Takahashi. 1991. The acquisition of echo questions. *University of Massachusetts Occasional Papers in Linguistics*, 17(1):11.

Koji Tanaka, Junya Takayama, and Yuki Arase. 2019. Dialogue-act prediction of future responses based on conversation history. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 197–202.

Deborah Tannen. 2007. *Talking voices: Repetition, dialogue, and imagery in conversational discourse*, volume 26. Cambridge University Press.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Edy Veneziano, Hermine Sinclair, and Ioanna Berthoud. 1990. From one word to two words: repetition patterns on the way to structured speech. *Journal of child language*, 17(3):633–650.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.

David Wood, Jerome S Bruner, and Gail Ross. 1976. The role of tutoring in problem solving. *Journal of child psychology and psychiatry*, 17(2):89–100.

Charles Yang, Stephen Crain, Robert C Berwick, Noam Chomsky, and Johan J Bolhuis. 2017. The growth of language: Universal grammar, experience, and principles of computation. *Neuroscience & Biobehavioral Reviews*, 81:103–119.

Victor H Yngve. 1970. On getting a word in edgewise. In *Papers from the sixth regional meeting Chicago Linguistic Society, April 16-18, 1970, Chicago Linguistic Society, Chicago*, pages 567–578.

Piotr Żelasko, Raghavendra Pappagari, and Najim Dehak. 2021. What helps transformers recognize conversational structure? importance of context, punctuation, and labels in dialog act recognition. *Transactions of the Association for Computational Linguistics*, 9:1163–1179.

Hao Zhu, Yonatan Bisk, and Graham Neubig. 2022. Language learning from communicative goals and linguistic input. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Language Resource References

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, 44:978–990.

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

A. Feedback Definitions & Annotation Guidelines

A.1. Definitions

Modelling *Modelling* indicates an adult modelling how to answer a question. The adult initiates a question-and-answer sequence without any contribution from the child. This feedback type captures instances where an adult demonstrates question-answering behaviour, potentially serving as a model for the child.

Exemplifying *Exemplifying* refers to an instance 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.

Reformulating *Reformulating* 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.

Affirming *Affirming* consists of the caregiver providing feedback on a question-and-answer sequence initiated by the child. 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.

A.2. Extraction Rules

Feedback candidates are extracted using algorithms created to reflect the broad linguistic heuristics used when developing the annotation guidelines. Algorithms are outlined in 1, 2, 3 and 4. Preprocessing used SpaCy and NLTK for tokenization.

Algorithm 1 Modelling Extraction

Require: *list*(utterances u, speaker roles s)
  for each utterance pair p (u_1, u_2): **do**
    if speaker $s_1 = s_2$ and $s_1, s_2 \in Adult$ **then**
      if u_1 ends with ? **then**
        if *not*(u_2 ends with ?) **then**
          Extract p
        end if
      end if
    end if
  end for

Algorithm 2 Exemplifying Extraction

Require: *list*(utterances u, speaker roles s, vocabulary overlap vo)
  for each utterance triple t (u_1, u_2, u_3): **do**
    if speaker $s_1 \in Child$ **then**
      if speaker $s_2 = s_3$ and $s_2, s_3 \in Adult$ **then**
        if u_1 ends with ? **then**
          if u_2 ends with ? **then**
            if *not*(u_3 ends with ?) **then**
              get vo between (u_1, u_2)
              if $0 < vo < 1$ **then**
                Extract p
              end if
            end if
          end if
        end if
      end if
    end if
  end for

Algorithm 3 Reformulating Extraction

Require: *list*(utterances u, speaker roles s, vocabulary overlap vo)
  for each utterance pair p (u_1, u_2): **do**
    if speaker $s_1 \in Child$ and $s_2 \in Adult$ **then**
      if u_1 ends with ? **then**
        if u_2 ends with ? **then**
          get vo between (u_1, u_2)
          if $0 < vo < 1$ **then**
            Extract p
          end if
        end if
      end if
    end if
  end for

Algorithm 4 Affirming Extraction

Require: *list*(utterances u, speaker roles s, vocabulary overlap vo)
  for each utterance triple t (u_1, u_2, u_3): **do**
    if speaker $s_1 = s_2$ and $s_1, s_2 \in Child$ and speaker $s_3 \in Adult$ **then**
      if u_1 ends with ? **then**
        if *not*(u_2 and u_3 end with ?) **then**
          get vo between (u_2, u_3)
          if $0 < vo < 1$ **then**
            Extract p
          end if
        end if
      end if
    end if
  end for

A.3. Annotation Process and Guidelines

We provide annotators with a feedback candidate (which we extract from a dialogue if it matches a specific speaker interaction sequence, see A.2) with the 5 utterances preceding it in a dialogue. We provide a definition of the feedback category, and ask annotators to label whether this item should be labelled with that feedback category.

Initially, two linguists created the definitions and guidelines (second and third authors). They independently annotated and arrived at refined coding scheme. The other two authors (computational linguistics, NLP training, no formal linguistics background) annotated a final set to test the robustness of the guidelines. Cohen's kappa was used: the resulting agreement was substantial, nearing almost perfect for certain categories. Any non-agreement cases were referred to the trained linguists. The four authors of this paper worked as annotators and refined the guidelines and distinctions between categories in stages. Inter-annotator agreement was recorded at each stage. Details are provided in Table 5.

	M	E	R	A	Av.
JH + EE First Set	0.43	0.17	0.67	0.39	0.42
JH + EE Second Set	0.80	0.45	0.75	0.60	0.65
AS + JH,EE Agreed	0.68	0.96	0.91	0.80	0.83
MB + JH,EE Agreed	0.71	0.86	0.81	0.73	0.77
AS + MB	0.77	0.86	0.85	0.82	0.82

Table 5: Inter Annotator Agreement (IAA) scores measured with Cohen's kappa. M: Modelling, E: Exemplifying, C: Reformulating, A: Affirming. Av.: average across categories All final labels were arrived at via re-annotation of any ambiguous items, and any remaining items were resolved by discussion. We report the progress of agreement and refinement here.

Childes annotation formatting Within the CHILDES corpus, some sequences will contain unheard or muffled words which in the transcript look like the following:

```
Child: I think yyy ,  know ,  know know why ?
Child: they don't yyy yyy yyy ,  they don't bite
       Mommy ,  they don't bite yyy .
Adult: so the ,  so that they don't bite any of
       us .
```

We deal with these cases when calculating lexical overlap by excluding them from the utterance length.

B. Measures

B.1. Perplexity

We compute the perplexity values for each utterance in the annotated dataset using the pre-trained GPT-2 model (Radford et al., 2019). We hypothesise that badly-formed 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. We compute the perplexity of an utterance given the 5 previous utterances preceding it. These utterances, referred to as the *context utterances*, are concatenated with the current utterance to form the input sequence, excluding their speaker labels.

B.2. Age of Acquisition

We compute the average age of acquisition of all utterances in the feedback pair/triple using the age-of-acquisition ratings for 30,000 English words (Kuperman et al., 2012). With this, we take the age of acquisition of each word in an utterance and calculate the average of these values to give us the average age of acquisition of an utterance.

B.3. POS proportion

To calculate the proportion of different part of speech (POS) for each child utterance, we extract the tags and the count of each tag is calculated. The negation count also includes common contractions like "can't" and "won't", which are not explicit labelled as negation in the morphosyntactic (MOR) analysis of the CHILDES (MacWhinney, 2000) database. To account for this, we included words ending in *n't* as negation to ensure a more comprehensive measure. These counts are then normalized by the total number of POS tags in each utterance to provide the relative proportion for each POS category.

B.4. Structural Distance

As a method to capture the structural difference between a child question and an adult reformulation,

or a question and it's answer, regardless of speaker, we make use of the Levenshtein distance (Levenshtein et al., 1966) in terms of the part of speech tags between the two utterances. This is a common method for measuring the distance between two sequences, and has been used in previous work on child language acquisition (e.g., (Brodsky and Waterfall, 2007)). We use the implementation of the Levenshtein distance provided by the Python library 'distancia' [13]. This allows us to capture the structural similarity between the two utterances, regardless of the specific words used. We then analyse how this structural distance changes with the age of the child, and how it differs across different types of feedback.

C. Data

We extracted all dialogues that met our criteria from the childes corpus, and provide a breakdown of the resulting makeup of the data in Table 6.

Corpus	Start Age	End Age	Dialogues
Belfast	2;0	4;0	70
Bloom1970	1;8	3;1	25
Braunwald	1;5	3;11	173
Brown	1;6	4;0	147
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	271
Providence	1;0	4;0	349
Sachs	1;2	3;8	90
Snow	2;5	3;9	40
Suppes	1;11	3;3	51
Thomas	2;0	4;0	326
Weist	2;1	4;0	141

Table 6: Overview of the unique corpora included in the dataset.

D. Experimental Setup

We split the data from each feedback type into an 80-20 train test split, from a balanced set of positive and negative examples, sampling equivalent negative examples from our annotations to match the positive examples.

E. Logistic Regression Features

Features used for each feedback category are presented in the following:

Modelling This category includes two adult utterances `(A1, A2)`. The features used are:

- `A1_perp, A2_perp`: Perplexity of each utterance.
- `cosine_sim`: Cosine similarity between the two utterances.
- `perp_diff`: Difference in perplexity between the two utterances.
- `VO`: Vocabulary overlap between the utterances.

Exemplifying For utterance triples `(C, A1, A2)`, the following features are used:

- `C_perp`, `A1_perp`, `A2_perp`: Perplexity of each utterance.
- `cosine_sim1`, `cosine_sim2`: Cosine similarity between `C` and `A1`, and between `A1` and `A2`.
- `VO1`, `VO2`: Vocabulary overlap between `C` and `A1`, and between `A1` and `A2`.

Reformulating For utterance pairs `(C, A)`, the following features are used:

- `C_perp`, `A_perp`: Perplexity of the child's and adult's utterances.
- `cosine_sim`: Cosine similarity between `C` and `A`.
- `perp_diff`: Difference between `C_perp` and `A_perp`.
- `VO`: Vocabulary overlap between `C` and `A`.

Affirming For a triple of utterances `(C1, C2, A)`, the following features are used:

- `C1_perp`, `C2_perp`, `A_perp`: Perplexity of the first and second child utterances `C1`, `C2`, and the adult utterance `A`.
- `cosine_sim1`, `cosine_sim2`: Cosine similarity between `C1, C2` and `C2, A`.
- `VO1`, `VO2`: Vocabulary overlap between `(C1, C2)` and `(C2, A)`.

F. Prompting Design

We create 3 types of prompt template. We include include definitions given and positive and negative examples in Table 7. Models were instructed to respond with "yes" or "no". To ensure consistency, responses containing variations of 'yes' were mapped to 1 (positive label), while other responses, including ambiguous of unclear outputs, were mapped to 0 (negative label).

$Prompt_{def}$

```
You are an expert AI assistant designed to
recognise feedback in child-adult dialogue
excerpts. The category you are recognising
is [Feedback Name] Feedback.
Carefully pay attention to the definitions
provided below!

Definitions:
[Feedback Definition]

Consider the excerpt carefully, and determine
```

[13] Levenshtein implementation https://distancia.readthedocs.io/en/latest/Levenshtein.html

```
if it is Modelling Feedback based on the
definitions provided.
Respond only with Yes or No
Excerpt: {dialogue}
Answer:
```

$Prompt_{def+pos}$

```
You are an expert AI assistant designed to
recognise feedback in child-adult dialogue
excerpts. The category you are recognising
is [Feedback Name] Feedback.
Carefully pay attention to the definitions
provided below!

Definitions:
[Feedback Definition]

Below is a positive example of [Feedback Name]
Feedback:
[Feedback Example]

Consider the excerpt carefully, and determine
if it is
Modelling Feedback based on the definitions
provided.
Respond only with Yes or No
Excerpt: {dialogue}
Answer:
```

$Prompt_{def+pos+neg}$

```
You are an expert AI assistant designed to
recognise feedback in child-adult dialogue
excerpts. The category you are
recognising
is [Feedback Name] Feedback.
Carefully pay attention to the
definitions provided below!

Definitions:
[Feedback Definition]

Below is a positive example of [Feedback Name]
Feedback:
[Feedback Example]

Below is a negative example of [Feedback Name]
Feedback: [Feedback Example]

Consider the excerpt carefully, and determine
if it is Modelling Feedback based on the
definitions provided.
Respond only with Yes or No
Excerpt: {dialogue}
Answer:
```

G. Results

Table 8 shows performance for two additional instruction tuned LMs. These results are disappointing and far lower than llama the best performing of this method, thus we only included it in the main body of the paper. We include these results to demonstrate that a labelling task of this complexity is beyond the capability of the smaller LMs, it is likely that if we were to use a larger, 70b parameter model (out of compute scope for this project) that we would see a boost in performance.

G.1. Easy and Hard Sets

Results for easy and hard sets can be found in Figure 8. These show in line with our expectations and experience with annotation: that the highest predictive features are cosine similarity and VO between utterances for Reformulating and the first pair of utterances in Exemplifying, we observe a similar pattern for modelling and affirming, but to a lesser extent. We include examples of some common errors in Table 10.

Figure 8 shows the feature importance for the Logistic regression models, demonstrating the importance for between utterance coherence and cohesion, measured by lexical and semantic similarity.

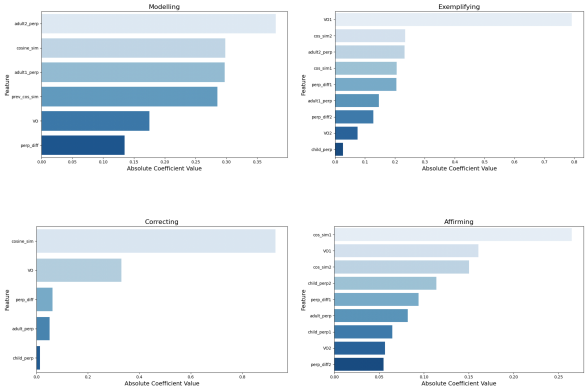

Figure 8: Feature importance for logistic regression.

H. Analysis

H.1. Incidence Across Development

Results of the significance tests across age bins are reported in Table 11. These tests support the significance labels in Figure 2.

H.2. Gold annotation analysis

With all our analyses excepting the feedback incidence over development, which we cannot directly compare, we compared trends observed in the full corpus against the original set of gold labels, observing similar relationships between the utterance length; lexical sophistication; POS proportion; lexical repetition; and structural repetition present in the feedback across development. We conducted this analysis as a sanity check of the results and analysis present in our work. Results can be found in Table 12, directionality differences are highlighted, we observe these for POS proportion and VO in Reformulation feedback. We report the incidence of feedback present in the gold dialogues to demonstrate that at a coarse scale we can observe the same ordering across feedback types with development, highlighting that without the scale of the automatic annotation, we would be far less able to conduct an analysis of this level of detail. We

Modelling	Exemplifying	Reformulating	Affirming
Definitions: - Modelling Feedback refers to a situation when an ADULT demonstrates how to answer a question to a CHILD. - ADULT initiates a question-and-answer sequence without any contribution from the CHILD. - If the ADULT **does not** answer their own question, the answer is **NO**.	**Definitions:** - Exemplifying Feedback refers to a situation where an ADULT rephrases of a CHILD'S question due to non-adult grammar. - The CHILD first asks a question, then ADULT repeats or reformulates the question. - The ADULT then provides an answer after rephrasing the CHILD'S question. - If the ADULT **partially repeats** the CHILD'S question, the answer is **NO**.	**Definitions:** - Correcting Feedback refers to a situation where an ADULT rephrases a CHILD'S question due to non-adult grammar. - The CHILD first asks a question, then ADULT rephrases the question. - The ADULTS's rephrasing HAS to provide a corrected version of the CHILD's utterance. - If the ADULT **partially repeats** the CHILD'S question, the answer is **NO**.	**Definitions:** - Affirming Feedback refers to a situation where an ADULT provides feedback on a question-and-answer sequence initiated by a CHILD. - The CHILD HAS to ask a question and then ANSWER it. - The ADULT then VALIDATES the CHILD'S answer. - If the CHILD **does not** answer their own question, or if the ADULT **does not** validate it, the answer is **NO**.
Positive example M: ADULT: are your hands clean now ? ADULT: oh they're pretty clean .	**Positive example E:** CHILD: a Anne's juice ? ADULT: where's Anne's juice ? ADULT: it's here	**Positive example C:** CHILD: where's moon ? ADULT: where's the moon ?	**Positive example A:** CHILD: who is that ? CHILD: this is Nathaniel's pancake pan . ADULT: that's Nathaniel's pancake pan that's right .
Negative example M: ADULT: can you narrow it down a little bit ? ADULT: bye bye pirates .	**Negative example E:** CHILD: the monster's growing all up our cars . ADULT: okay . ADULT: but I'll hafta make him first won't I ?	**Negative example C:** CHILD: more grape juice . ADULT: more ?	**Negative example A:** CHILD: where is Mama ? CHILD: if Laura raise the bathroom door to tell Jack that it's breakfast time and she's saying knock knock . ADULT: okay Laura .

Table 7: Prompt definitions of each category presented to the instruction-tuned LMs we investigate. Note: in our initial experiments we referred to Reformulating as *Correcting*, which we later refined to more closely reflect the nature of the feedback we focus on.

		M				E				R				A				Av.
		Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc
Llama3 8b-Instruct	$Prompt_{def}$	**0.57**	0.60	0.39	0.47	0.56	0.55	0.65	0.60	**0.69**	0.62	1.00	0.76	**0.64**	0.61	0.79	0.69	0.62
	$Prompt_{def+pos}$	0.48	0.40	0.09	0.14	0.66	0.61	0.90	0.72	0.66	0.59	1.00	0.74	0.64	0.61	0.77	0.68	0.61
	$Prompt_{def+pos+neg}$	0.48	0.44	0.17	0.25	**0.71**	0.64	0.98	0.77	0.67	0.60	1.00	0.75	0.63	0.59	0.86	0.70	0.62
Gemma 2b-Instruct	$Prompt_{def}$	0.50	0.50	1.00	0.67	0.49	0.49	0.98	0.66	0.50	0.50	1.00	0.67	0.55	0.52	1.00	0.69	0.51
	$Prompt_{def+pos}$	0.51	0.51	0.87	0.65	0.44	0.46	0.69	0.55	0.50	0.50	1.00	0.67	0.47	0.48	0.86	0.62	0.48
	$Prompt_{def+pos+neg}$	0.52	0.51	0.91	0.66	0.47	0.48	0.94	0.64	0.50	0.50	1.00	0.67	0.48	0.48	0.74	0.59	0.49
Falcon 7b-Instruct	$Prompt_{def}$	0.50	0.50	1.00	0.67	0.48	0.49	0.96	0.65	0.48	0.48	0.48	0.48	0.50	0.50	1.00	0.67	0.49
	$Prompt_{def+pos}$	0.50	0.50	1.00	0.67	0.49	0.49	0.98	0.66	0.36	0.41	0.62	0.49	0.50	0.50	1.00	0.67	0.46
	$Prompt_{def+pos+neg}$	0.50	0.50	1.00	0.67	0.50	0.50	1.00	0.67	0.43	0.46	0.83	0.59	0.50	0.50	1.00	0.67	0.48

Table 8: Evaluation results across feedback types with average accuracy. M: Modelling. E: Exemplifying. R: Reformulating. A: Affirming. Acc: Accuracy. Prec: Precision. Rec: Recall. F1: F1-score. **Bold** and underline indicates the best overall model accuracy for each feedback category.

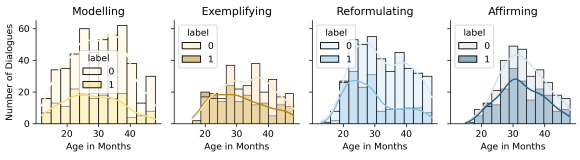

Figure 9: Distribution of annotated dialogues. labels compare quantity of dialogues containing positive (1) vs false (0) examples of feedback.

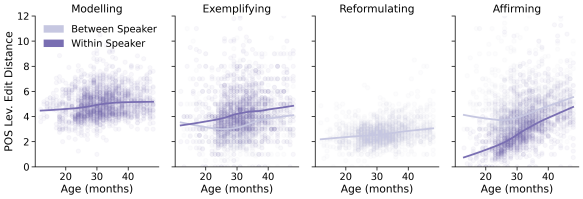

Figure 10: POS Levenshtein Edit Distance between and within speaker across feedback categories.

include the distribution across feedback types for completeness in Figure 9.

H.3. Syntactic Repetition

Figure 10 shows the Levenshtein edit distance between and within speaker of POS tags, a measure of syntactic repetition. Complementary to the Vocabulary overlap analysis in Figure 6, this demonstrates the child's progression towards adult-like syntactic relationship between questions and answers.

		M					E					R					A				
		Acc	Prec	Rec	F1	ΔF1	Acc	Prec	Rec	F1	ΔF1	Acc	Prec	Rec	F1	ΔF1	Acc	Prec	Rec	F1	ΔF1
BERT	Easy	-	-	-	-	-	0.85	0.80	0.92	0.85	+0.06	**0.80**	0.74	0.93	0.82	−0.15	0.78	0.77	0.79	0.78	−0.19
	Hard	-	-	-	-		**0.89**	0.83	1.00	0.91		0.60	0.50	1.0	0.67		0.61	0.63	0.56	0.59	
BERT + C	Easy	**0.78**	0.83	0.69	0.75	−0.25	0.78	0.73	0.87	0.80	+0.02	0.77	0.78	0.76	0.77	+0.23	0.76	0.75	0.79	0.77	−0.27
	Hard	0.60	0.75	0.38	0.50		0.78	0.75	0.90	0.82		1.0	1.0	1.0	1.0		0.44	0.45	0.56	0.50	
MBERT	Easy	-	-	-	-	-	0.83	0.78	0.92	0.84	+0.03	0.79	0.80	0.78	0.79	−0.12	**0.79**	0.72	0.97	0.83	−0.08
	Hard	-	-	-	-		0.83	0.77	1.0	0.87		0.80	1.0	0.50	0.67		0.67	0.60	1.0	0.75	
MBERT + C	Easy	0.58	0.69	0.26	0.34	−0.34	0.65	0.59	0.97	0.73	+0.05	0.72	0.96	0.47	0.63	+0.04	0.75	0.72	0.82	0.77	−0.10
	Hard	0.47	0	0	0		0.72	0.69	0.90	0.78		0.80	1.0	0.50	0.67		0.61	0.58	0.78	0.67	

Table 9: Evaluation results across feedback types for easy and hard sets. M: Modelling. E: Exemplifying. R: Reformulating. A: Affirming. Acc: Accuracy. Prec: Precision. Rec: Recall. F1: F1-score. ΔF1: difference in F1 (Hard − Easy); gray indicates higher performance on *harder*—more ambiguous to a human rater—items. **Bold** and underline indicates the best overall model accuracy for each feedback category. + C is plus context.

Modelling	Exemplifying	Reformulating	Affirming
Positive example: - ADULT: they get cross don't they ? - ADULT: they get sick of it .	**Negative example:** CHILD: where she gone ? ADULT: where's she gone ? ADULT: has she gone shopping .	**Negative example:** CHILD: where's green ? ADULT: green what ?	**Negative example:** CHILD: where Percy ? CHILD: in Percy . ADULT: here's Percy .
Difficulty level: Easy Misclassified as negative by BERT, Logistic Regression and Llama3 8b-Instruct.	**Difficulty level: Hard** Misclassified as positive by all models	**Difficulty level: Easy** Misclassified as positive by BERT, Llama3 8b-Instruct, Gemma 2b-Instruct and Falcon 7b-Instruct	**Difficulty level: Hard** Misclassified as positive by all models

Table 10: Examples of Incorrectly classified examples from the test set

Comparison	M_1	M_2	t	$p_{corrected}$
Modelling				
<20 vs. <30	0.149	0.090	7.659	<0.001
<30 vs. <40	0.090	0.075	7.852	<0.001
<40 vs. <50	0.075	0.062	5.175	<0.001
Exemplifying				
<20 vs. <30	0.004	0.006	-5.616	<0.001
<30 vs. <40	0.006	0.006	2.014	0.133
<40 vs. <50	0.006	0.005	0.218	0.828
Reformulating				
<20 vs. <30	0.020	0.025	-2.735	0.029
<30 vs. <40	0.025	0.022	3.022	0.015
<40 vs. <50	0.022	0.018	4.626	<0.001
Affirming				
<20 vs. <30	0.007	0.016	-8.735	<0.001
<30 vs. <40	0.016	0.017	-1.550	0.243
<40 vs. <50	0.017	0.020	-2.826	0.025

Table 11: Incidence across development. Independent t-test results, adjusted with Bonferroni corrections for multiple tests. gray indicates insignificance.

Property	**M**	**E**	**R**	**A**
MLU				
child	-	0.333	0.550	0.402
adult	0.209	0.213	0.195	0.441
AoA				
child	-	-0.043	0.075	-0.008
adult	0.024	-0.098	0.003	0.026
POS Prop.				
Verb	-	0.095	-0.055	0.125
Adj	-	0.009	0.235	0.060
Noun	-	-0.063	-0.118	-0.008
Aux	-	-0.038	0.041	-0.038
Det	-	0.064	0.112	0.049
Neg	-	0.047	0.230	0.157
VO				
between	-	-0.072	0.216	-0.031
within	-0.079	0.061	-	0.028
Struct. Dist.				
between	-	0.148	0.222	0.159
within	0.196	0.064	-	0.374

Table 12: Gold annotations: Pearson correlation coefficients (r). Insignificant values ($p > 0.05$) in light gray. Different relationships from the scaled up annotation (Table 4 highlighted in teal. Mean length of utterance (MLU), lexical sophistication (AoA), POS proportion, vocabulary overlap (VO) and Structural Distance (levenshtein edit distance between- or within-speaker utterances) with respect to child age in months, across feedback types and speakers.