Beyond Length: Context-Aware Expansion and Independence as Developmentally Sensitive Evaluation in Child Utterances

Jiyun Chun, Eric Fosler-Lussier, Michael White, Andrew Perrault
The Ohio State University
{chun.203,fosler-lussier.1,white.1240,perrault.17}@osu.edu

Abstract

Evaluating the quality of children's utterances in adult-child dialogue remains challenging due to insufficient context-sensitive metrics. Common proxies such as Mean Length of Utterance (MLU), lexical diversity (vocd-D), and readability indices (Flesch-Kincaid Grade Level, Gunning Fog Index) are dominated by length and ignore conversational context, missing aspects of response quality such as reasoning depth, topic maintenance, and discourse planning. We introduce an LLM-as-a-judge framework that first classifies the **Previous Adult Utterance Type** and then scores the child's response along two axes: **Expansion** (contextual elaboration and inferential depth) and **Independence** (the child's contribution to advancing the discourse). These axes reflect fundamental dimensions in child language development, where Expansion captures elaboration, clause combining, and causal and contrastive connectives. Independence captures initiative, topic control, decreasing reliance on adult scaffolding through growing self-regulation, and audience design. We establish developmental validity by showing age-related patterns and demonstrate predictive value by improving age estimation over common baselines. We further confirm semantic sensitivity by detecting differences tied to discourse relations. Our metrics align with human judgments, enabling large-scale evaluation. This shifts child utterance assessment from simply measuring length to evaluating how meaningfully the child's speech contributes to and advances the conversation within its context.[1]

1 Introduction

Child-facing AI systems are increasingly deployed in settings such as educational chatbots, museum guides, and language learning tools (Ruan et al., 2019; Labadze et al., 2023; Beelen et al., 2024; Du and Daniel, 2024). In these applications, designers and researchers need ways to evaluate child utterances beyond surface length, capturing how a child's turn fits the local question-answer relation and how it advances discourse. Beyond age appropriateness, our broader aim is to develop context-aware metrics that capture developmentally relevant signals in real child-adult conversations.

Evaluating children's language development in dialogue remains challenging. Widely used proxies such as Mean Length of Utterance (MLU; Brown 1973), type-token ratios (TTR; Templin 1975), and readability formulas including Flesch-Kincaid Grade Level (Kincaid et al., 1975) and Gunning Fog Index (Gunning, 1952) were designed for written text rather than spontaneous child-adult conversation. Length-based measures capture quantity but often fail to distinguish long yet content-poor utterances from brief yet well-reasoned ones. TTR variants are sensitive to sample size (Treffers-Daller et al., 2016; Richards, 1987). vocd-D reduces but does not remove this sensitivity, remaining unreliable on very short turns (McCarthy and Jarvis, 2007; Ratner et al., 2024). Readability indices depend on sentence boundaries and syllable counts, which makes them brittle in the presence of disfluencies, fragments, and backchannels that are common in child speech (Redish, 2000; Greenfield and McNamara, 2008; Macwhinney, 2000; Tran et al., 2020). More broadly, these proxies tend to be context insensitive because they ignore how a child's turn relates to the preceding adult prompt such as question type and discourse function. They also correlate strongly with length, which encourages a length as proxy interpretation where verbosity is mistaken for competence. We therefore need evaluation measures that are sensitive to semantic and pragmatic contributions in context and that can disentangle substantive development from mere increases in utterance length.

A more fundamental limitation is the lack of con-

---

[1] https://github.com/Amyyyyeah/Beyond-Length

text awareness. In child-adult dialogue, a child's response is tightly linked to the immediately preceding adult turn (contingency; Bloom et al. 1976; Rowe 2012; Callanan et al. 2020; Pagmar et al. 2022; Zohrabi et al. 2014). Two responses of similar length may differ greatly in inferential depth and alignment to the question. For example, to "What did you do at school today?", "I don't really remember. We just hung out at recess, nothing special, I guess" is vague and weakly aligned, whereas "We started a bean-germination experiment, labeled our pots, and we'll check for sprouting." conveys concrete actions, temporal structure, and a plan, thereby expanding the prior context. Metrics that ignore question–answer relations and discourse functions (e.g., topic maintenance, repair, joint planning) tend to underestimate such developmental signals (Rowe et al., 2017; Clark, 2020; Matthews et al., 2018).

Psycholinguistics and developmental linguistics have established concepts of conversational development: contingency (contextually appropriate responses), elaboration (addition of details and reasoning), scaffolding (utterance facilitation), and topic maintenance (Bloom et al., 1976; Fivush et al., 2006; Ninio and Snow, 1996). These frameworks can be summarized along two axes: children's ability to independently advance discourse and their capacity to expand upon previous turns' meaning (Salomo et al., 2013; Luo et al., 2022). Nevertheless, quantitative metrics for automatically measuring these concepts at scale have been absent. Prior work mainly relied on rule-based pipelines (Macwhinney, 2000; Miller et al., 2012) or supervised classifiers (Lin et al., 2023; Stolcke et al., 2000). Rule-based pipelines were brittle to transcription artifacts such as disfluencies and incomplete clauses and transferred poorly across datasets or recording conventions (Eisenberg et al., 2001; Guo and Eisenberg, 2015). Supervised classifiers captured local surface features but struggled to condition on the full question-answer relation and discourse functions without heavy feature engineering, and they required retraining whenever the evaluation rubric changed. Recent advances in large language models (LLMs) change this feasibility boundary. LLMs can follow natural language rubrics, read context, and map free text to interpretable ordinal ratings without task-specific fine-tuning (Zheng et al., 2023; Gu et al., 2025; Liu et al., 2023).

We adopt an LLM-as-a-judge framework that classifies adult prompts by **Previous Adult Utterance Type** (*PT*: question and non-question subtypes) and evaluates the corresponding child responses on two theoretically-motivated dimensions: **Expansion** (contextual elaboration and inferential depth) and **Independence** (autonomous contribution to discourse). To our knowledge, this study is the first to enable large-scale, context-conditioned scoring of child utterances and to demonstrate both developmental signal and predictive utility while controlling for utterance length and *PT*. Empirically, our scores show systematic variation across *PT* and remain significantly associated with age even after controlling for length and *PT*. This overcomes limitations of length-dominated proxies, which fail to distinguish long-but-thin from short-but-well-reasoned responses. Moreover, at fixed utterance length, our metrics capture differences in reasoning depth and contextual fit, shifting evaluation from quantity to quality of discourse contribution. We further show that our approach provides effective predictive features for age and exhibits strong agreement with human evaluation, offering rubric alignment and interpretability that length/syllable-based metrics lack. All analyses in this paper are text-based, evaluating orthographic transcripts rather than audio. Incorporating acoustic and prosodic features is left for future work.

2 Related Work

Child language research has largely relied on large-scale transcript corpora such as CHILDES (Macwhinney, 2000) and on standard sentence- or token-level metrics. Representative measures include Mean Length of Utterance (MLU; Ezeizabarrena and Fernández 2018; Guo and Eisenberg 2015; Rice et al. 2010), lexical diversity (TTR; Hess et al. 1986, 1989, vocd-D; Yang et al. 2022; deBoer 2014), and readability indices (Flesch-Kincaid; Murphy et al. 2022; Sharratt et al. 2025, Gunning Fog; Louw et al. 2025), with some studies supplementing clause- or phrase-based complexity indicators (Nippold et al., 2005). However, these metrics suffer from three key limitations. First, length and complexity are confounded, making interpretation difficult. Second, stable estimation requires lengthy texts. Third, they fail to directly account for discourse context, such as the type of preceding adult question. We propose discourse-level metrics that separately measure content Expansion (E) and Independence (I) in child utter-

ances. These definitions align with established strands of child development research: *E* parallels discourse elaboration (Evers-Vermeul and Sanders, 2011; Veen, 2011; Thwaite, 2019; Toppelberg and Collins, 2004; McCabe, 1996; Nippold et al., 2005; Kallay and Redford, 2020) in usage-based and narrative work (clause density, connective use, and detail), whereas *I* tracks children's development from adult-dependent responses to self-initiated contributions, a developmental trajectory predicted by scaffolding and topic-management research (Toppelberg and Collins, 2004; Abbot-Smith et al., 2023; VanDam et al., 2022; Masek et al., 2021; Snow, 1977; Kelly et al., 2022; Alduais et al., 2022; Leech et al., 2013; Chouinard, 2007).

Child language development extends beyond mere vocabulary accumulation to encompass the ability to expand and organize information within discourse and to advance conversation proactively. Adult interaction patterns, question types, and feedback strategies exert strong influence on the form and function of child utterances (Toppelberg and Collins, 2004; Rowe, 2012; Callanan et al., 2020; Rowe and Snow, 2019; Salomo et al., 2013). For instance, yes/no and display questions (where asker already knows the answer) tend to elicit brief responses (Wright, 2016), whereas referential and open questions, along with utterances incorporating contrast or causal markers and initiative-eliciting moves, promote longer explanations and the introduction of new information (Thorson et al., 2024; Raising Children Network, 2025). Although existing research has qualitatively noted that adult utterances shape children's immediate responses, no prior work has simultaneously: (1) quantified contextual effects at the utterance level, (2) disentangled developmental signals from length, and (3) demonstrated that these signals remain valid in predictive tasks such as age estimation.

3 Methods

Datasets We used corpora from the Child Language Data Exchange System (CHILDES (MacWhinney, 2014)), a large global archive of manually transcribed child–caregiver conversations ranging from birth to age 13. We focused on the English North American (Eng-NA) subcorpus and excluded sessions involving clinical populations and non-spontaneous interactions such as singing, verse recitation, and alphabet or number reading. As detailed in Appendix C, our study aggregates data from 20 distinct corpora (e.g., Brown, Clark) within CHILDES, comprising over 360K utterances and exhibiting high variance in speakers and recording conditions. Our target age range is 2–10 years, as this period captures the most active phase of syntactic and pragmatic development.

Previous Adult Utterance Type Research on child language shows that children's turns are strongly conditioned by the discourse function of the preceding adult utterance, which can be a question, a non-question, or a special case. We standardize the Previous Adult Utterance Type (*PT*) to capture how the pragmatic demand of the immediately preceding adult move shapes children's responses. Caregivers use different kinds of questions to scaffold participation, narrowing or widening the response space. Closed questions (e.g., yes/no, repetition) elicit brief forms, while open questions invite elaboration and child-initiated contributions. As shown in Table 1, we classify adult utterances into questions and non-questions, each with further subdivisions. We use Gemini 2.5 Pro (Comanici et al., 2025) for automated annotation based on these *PT* types. Detailed definitions and supporting literature for each *PT* category are provided in Appendix A.

Expansion To measure developmental sophistication beyond mere length, Expansion (*E*) quantifies contextual elaboration and inferential depth, that is, how narratively and logically organized a child utterance is, with scores from 1 to 10. This definition aligns with established strands of discourse elaboration research examining clause density, connective use, and narrative detail (Evers-Vermeul and Sanders, 2011; Veen, 2011; Thwaite, 2019; Toppelberg and Collins, 2004; McCabe, 1996; Nippold et al., 2005; Kallay and Redford, 2020). Unlike length-based proxies that fail to distinguish long-but-thin from short-but-well-reasoned responses, *E* rewards logical organization and discourse linkage. In brief, the core scale progresses from single word (1), single clause (4), coordinating linkage (6), single causal relation (7), temporally ordered sequence (8), complex structure with subordination (9), to integrated narrative across multiple times or perspectives (10). Judgments rely on coordinating or subordinating connections, on explicit markers of cause, condition, or contrast, and on temporal or logical ordering.[2]

---

[2]Detailed scoring criteria and rubric definitions for Expansion and Independence are provided in Appendix B.

Independence Beyond responding appropriately, children gradually learn to manage and direct conversation independently. Independence (*I*) measures the extent to which the child advances and regulates the interaction without relying on the adult context with scores from 1 to 10. *I* quantifies children's progression toward autonomous conversational control, aligning with findings on scaffolding reduction and self-directed topic management (Toppelberg and Collins, 2004; Abbot-Smith et al., 2023; VanDam et al., 2022; Masek et al., 2021; Snow, 1977; Kelly et al., 2022; Alduais et al., 2022; Leech et al., 2013; Chouinard, 2007). Providing a correct answer alone does not increase independence. We prioritize interactional functions such as unsolicited justification, self-initiated exploratory questions, topic management, and listener-oriented regulation. Scores progress from echoing and minimal response (1-4), through self-initiated questions and justifications (5-7), to multi-act regulation and conversational leadership (8-10).[2]

4 Results and Analysis

We first examine the distribution of these dimensions and demonstrate that they capture distinct developmental mechanisms rather than merely tracking utterance length. We then assess their developmental validity through an age prediction task, where we also compare against open-weight models. Additionally, we show that our dimensions can differentiate contextual nuances even when utterance lengths are similar. Finally, we validate that our automated annotations align with human judgments via human evaluation (Section 4.5).

4.1 Distributions of Expansion and Independence

We examine how Expansion (*E*) and Independence (*I*) distribute across question and non-question types. Table 1 presents descriptive statistics and the number of utterances *N* for each previous adult utterance type (*PT*). The reported statistics (mean, standard deviation, median, inter-quartile range) are computed over the child utterances that respond to those adult turns. As expected, open-ended questions tend to elicit richer child elaboration. The mean of Expansion is highest for *Elaboration*, while closed-question types such as *Yes/No* and *Display* exhibit lower means. For Independence, non-question turns show higher central tendency than question turns. In our data, non-question types

have higher means and medians (4.81-5.16; 5) than question types (4.21-4.76; 4). Conversely, constrained question types limit opportunities for autonomous child talk. This constraint is reflected in *I*, which shows narrow dispersion for *Display* and *Choice* questions (IQR=0). In contrast, open-ended *Elaboration* questions show the largest spread in *E*, reflecting greater response variability.

4.2 Developmental Validity

Do our metrics diagnose the kinds of changes theoretically expected with age, thereby establishing developmental validity? *E* is intended to reflect development that is realized through lengthened and better-organized discourse (multi-clause integration, causal/temporal linking). By contrast, *I* is designed to capture context-sensitive discourse management (initiative, clarification, regulation of topic flow) that can improve even when responses are not longer. Because chronological age co-varies with verbosity, naive age effects can collapse to longer utterances rather than genuine strategy changes. We therefore separate length-carried effects from length-independent ones and test whether *E/I* behave in line with these theoretical roles. All analyses in this section use *PT*, *E*, and *I* scores labeled by `Gemini 2.5 Pro`.

Setup As baseline comparators, we include Mean Length of Utterance (MLU; Brown 1973), lexical diversity (vocd-D; McKee et al. 2000), and readability indices (Flesch-Kincaid Grade Level; Kincaid et al. 1975, Gunning Fog Index; Gunning 1952). Since age and utterance length tend to increase together, the association between age and any outcome can collapse to a pure length effect unless length is separated. Yet length is necessary but not sufficient for development. Even at the same length, children can differ in discourse strategy (reasoning, clarification, topic management) and in their sensitivity to *PT*. Length-based metrics and readability formulas conflate verbosity with strategy, lack context sensitivity, and become unstable when length distributions differ across children.

To avoid conflating verbosity with development, we decompose length into within-child and between-child components, and adopt a two-model design that separates effects carried by length from effects that remain once length is held constant. Accordingly, we estimate a total-effect model (Equation 1) to assess the aggregate age association and a direct-effect model (Equation 2) to assess what

PT	Type	N	Expansion				Independence			
			Mean	SD	Median	IQR	Mean	SD	Median	IQR
Question	Yes/No	25883	2.64	2.06	2	3	4.39	1.63	4	2
	Referential	16924	3.60	2.20	4	3	4.46	1.42	4	1
	Clarification Request	3586	3.53	2.09	3	2	4.76	1.58	4	1
	Display	11893	2.68	1.77	2	3	4.21	1.17	4	0
	Confirmation Check	11438	2.85	2.14	2	3	4.52	1.77	4	2
	Choice	2012	3.11	2.06	3	3	4.36	1.20	4	0
	Elaboration	6148	4.46	2.46	4	3	4.66	1.42	4	1
Non-Question	Topic Introduction	19189	3.25	2.02	3	3	4.94	1.91	5	2
	Imperative/Request	14759	3.18	2.01	3	3	4.81	1.91	5	1
	Feedback/Evaluation	39235	3.75	2.19	4	3	5.16	1.89	5	2
	Self-Talk/Thinking Aloud	1815	3.38	2.05	3	2	5.10	1.93	5	2

Table 1: Summary of Expansion and Independence scores by *PT* (all annotated by Gemini 2.5 Pro). N indicates the number of adult turns.

remains once length is held constant.

Equations 1,2 define the shared linear predictor η. For *E/I*, we fit proportional odds cumulative link mixed models **CLMMs** and report odds ratios e^{β}. For baselines, we use linear mixed models **LMMs** and report effects as coefficients β.[3]

$$Y = \beta_0 + \beta_{\text{age}} \text{Age}_z + \beta_{\text{PT}}^{\top} PT + u_{\text{child}} + \varepsilon \quad (1)$$

$$Y = \beta_0 + \beta_{\text{age}} \text{Age}_z + \beta_w L_z^{(w)} + \beta_b L_z^{(b)} + \boldsymbol{\beta}_{\text{PT}}^{\top} PT + u_{\text{child}} + \varepsilon \quad (2)$$

We use β to denote regression coefficients and Y for the outcome variable (metric). Age_z is standardized age, and *PT* encodes Previous Adult Utterance Type as fixed effects. We decompose utterance length into two z-scored components computed within child×PT cells. u_{child} is a child-level random intercept and ε is the residual error. Intuitively, the **within-child** component $L_z^{(w)}$ captures how long the current utterance is relative to that same child's typical length under the same *PT* (turn-level deviation from the child–*PT* mean). It lets us test prompt-contingent, discourse-level adjustments at a fixed habitual length. By contrast, the **between-child** component $L_z^{(b)}$ summarizes the child's habitual verbosity (the child–*PT* mean length itself), reflecting how talkative that child is compared to peers under the same *PT*. Separating these two components prevents verbosity from being mistaken for development.

Analysis MLU indicates typical utterance length, defined as the average number of words (or morphemes) per utterance computed over multiple turns. In Table 2, children produce longer utterances as they get older, so in the total-effect model, MLU increases with age with a small slope (β=0.355). When we include length in the model to separate the influence of length itself, the age effect disappears and only self-prediction remains. Because MLU is length by definition, it captures the pattern that speech becomes longer with age but it does not isolate differences in strategy or context beyond length. The lexical diversity measure vocd-D appears to increase strongly with age at first glance (Total=16.07, Direct=17.00). However, the interpretation changes once we split length in two ways. vocd-D responds little to momentary within-child changes in length relative to a child's own mean (≈ 0), while it responds negatively to differences in average length between children (-3.22). This means that children who are on average more verbose tend to receive lower diversity estimates, so when length distributions differ across groups the age-diversity relationship can be distorted. In short, vocd-D is insensitive to whether a child happens to speak longer in a given moment and tends to be biased depending on how long that child speaks on average. The readability indices FKGL and GFI show little response to age. Instead, scores increase for children with longer average utterances (FKGL=2.24, GFI=1.12). These indices therefore act mainly as proxies for average length rather than reflecting strategy or context. Expansion is intended to capture development that is realized through length. Consistent with this intention, the total-effect model shows

---

[3]We fit proportional **CLMMs** ($P(Y \leq r) = \text{logit}^{-1}(\theta_r - \eta)$) for *E/I* and **LMMs** ($Y = \eta + \varepsilon$) for baselines.

Baselines (LMMs)	Model	Age Effect (β)	Length Effect (β)
Mean Length of Utterance (MLU)	Total	$\beta_{age} = 0.355$	–
	Direct	–	self-prediction
Vocabulary Diversity–D (vocd-D)	Total	$\beta_{age} = 16.07$	–
	Direct	$\beta_{age} = 17.00$	$\beta_w \approx 0$ (n.s.), $\beta_b = -3.22$
Flesch-Kincaid Grade Level	Total	≈ 0	–
	Direct	≈ 0	β_w n/a, $\beta_b = 2.24$
Gunning Fog Index	Total	≈ 0	–
	Direct	≈ 0	β_w n/a, $\beta_b = 1.12$
Our metrics (CLMMs)	**Model**	**Age Effect (OR)**	**Length Effect (OR)**
Expansion	Total	OR = 1.28 ($\beta_{age} = 0.248$)	–
	Direct	OR = 0.99 (n.s.)	$OR_w = 10.36$, $OR_b = 4.27$
Independence	Total	OR = 1.05 ($\beta_{age} = 0.053$)	–
	Direct	OR = 0.90 ($\beta_{age} = -0.107$)	$OR_w = 2.71$, $OR_b = 1.76$

Table 2: Baselines (LMM) vs. our metrics (CLMM) under Total vs. Direct specifications. Analyses use *PT*, *E*, and *I* labeled by Gemini 2.5 Pro. LMM reports β, CLMM reports odds ratios (OR). n.s. = $p>0.05$ (not significant).

that Expansion increases with age (1.28). In the direct-effect model where length is separated, the age effect is not significant (0.99), while the length components are very large. A child's score rises sharply when they speak longer than usual on a given turn (10.36), and children who are on average more verbose also tend to score higher (4.27). In other words, increases in Expansion are realized almost entirely through becoming longer, and children who are on average more verbose also tend to obtain higher *E* scores. Independence is designed to capture context and strategy beyond length. Although scores increase with length (2.71, 1.76), in the direct-effect model, the residual effect of age is below 1 after controlling for length and *PT* (0.90). This means that at the same length and the same *PT* condition, an older child can receive a lower *I* score. This aligns with the scoring design of Independence, where the upper levels from 5 to 10 impose no minimum length requirement and instead assess discourse functions such as managing the flow of conversation or shifting topics. Thus, it is not how long the child speaks but how the child responds to the prompt and which discourse move the child selects that determines the score. The fact that a negative residual age slope remains even after holding length and *PT* fixed shows that differences in strategy and context sensitivity exist among children even at the same length.

4.3 Age Prediction

This section validates whether the developmental pathway decomposition confirmed in Section 4.2 generalizes to age prediction, not merely providing explanatory power. We assess whether our metrics provide additional predictive power beyond baseline metrics (MLU, vocd-D, FKGL, GFI), demonstrating that the developmental signal carries predictive validity.

Setup All experiments use only child utterances as input, with the target variable being the child's age in decimal years (e.g., 4 years 7 months = 4.58 years). We employed GroupKFold (k=5) to ensure all utterances from the same child remain in the same fold, preventing data leakage. We report mean absolute error, mean squared error, coefficient of determination R^2, and 95% confidence intervals. FKGL, GFI, and vocd-D, which require sufficient text length for stable estimation, were computed by concatenating all utterances for each child. We evaluate both with and without *PT* features. For our metrics, which are ordinal and distributional by design, we summarize E/I per child using Low/Medium/High ratios (1-3/4-7/8-10), and also compute *PT*-wise macro means and standard deviations to reflect context sensitivity (details in Appendix E). We also include an LLM baseline using open-weight model `Mixtral-8x7B-Instruct-v0.1` (chosen for its efficient MoE architecture and 32K context) and closed-weight model `Gemini 2.5 Pro` that directly predict age from child utterances only. Additionally, `Mixtral` labels PT, Expansion, and Independence features for linear/non-linear models.

Results Table 3 shows that E/I features outperform all baselines under both Linear Regression (LR) and Gradient Boosting (GB). Under LR, the

Type	Feature set	Mixtral-8x7B-Inst.						Gemini 2.5 Pro					
		MAE			**MSE**			**MAE**			**MSE**		
LLM	Child utterances	1.30			2.75			0.51	1.03			2.25	0.51

Type	Feature set	**Linear Regression**			**Gradient Boosting**			**Linear Regression**			**Gradient Boosting**		
		MAE	MSE	R^2	MAE	MSE	R^2	MAE	MSE	R^2	MAE	MSE	R^2
	MLU	1.53	3.84	0.17	1.52	4.25	0.07	1.42	3.47	0.23	1.55	4.10	0.09
	MLU + *PT*	1.51	3.78	0.18	1.50	4.15	0.09	1.40	3.41	0.25	1.50	3.91	0.14
	vocd-D	1.76	4.58	0.01	1.81	4.87	-0.06	1.75	4.55	0.00	1.78	4.66	-0.03
	vocd-D + *PT*	1.72	4.47	0.03	1.76	4.63	0.00	1.71	4.41	0.03	1.70	4.35	0.04
Baselines	FKGL	1.76	4.55	0.02	1.73	4.56	0.01	1.72	4.40	0.03	1.67	4.27	0.06
	FKGL + *PT*	1.73	4.46	0.03	1.71	4.47	0.03	1.69	4.31	0.05	1.63	4.13	0.09
	GFI	1.66	4.22	0.09	1.52	3.80	0.18	1.57	3.92	0.14	1.43	3.43	0.25
	GFI + *PT*	1.64	4.17	0.10	1.52	3.75	0.19	1.56	3.87	0.15	1.40	3.31	0.27
	E + I	1.20	2.75	0.40	1.23	2.97	0.36	1.14	2.41	0.47	1.05	2.09	0.54
	E + I + PT	**1.19***	**2.71***	**0.41***	**1.22***	**2.89***	**0.37***	**1.13***	**2.38***	**0.48***	**1.04***	**2.06***	**0.55***
Ours	*E*	1.20	2.75	0.40	1.23	2.97	0.36	1.22	2.76	0.39	1.14	2.47	0.46
	E + PT	1.20	2.73	0.41	1.23	2.97	0.36	1.21	2.71	0.40	1.12	2.42	0.47
	I	1.34	3.17	0.31	1.35	3.38	0.26	1.24	2.76	0.39	1.33	3.03	0.33
	I + PT	1.33	3.13	0.32	1.34	3.33	0.27	1.23	2.72	0.40	1.33	3.03	0.33

Table 3: Age prediction with child-grouped 5-fold CV. (*) indicates the best performance. 95% confidence intervals (CIs) are reported in Appendix Tables 9 and 10.

best baseline is MLU with *PT* (1.40), whereas Ours (*E + I + PT*) attains a 19.3% MAE reduction. Under GB, the reduction is even larger at 25.7% compared to the best baseline GFI with *PT*. Although Section 4.2 shows that vocd-D increases with age ($\beta \approx$16-17), the actual age ranges overlap widely at the same vocd-D value, yielding low R^2 and the largest MAE in prediction. This arises because vocd-D exhibits $\beta_b < 0$ (Table 2), reflecting its sensitivity to habitual verbosity at the between-child level. We also observe sizeable gains in association with age, as R^2 improves by +0.23 in Linear (0.25→0.48) and +0.28 in GB (0.27→0.55) over the strongest baselines. While adding *PT* improves all feature sets, the gain is comparatively smaller for our *E/I* features than for traditional baselines, suggesting that our summary statistics already capture *PT*-specific effects. Analyzing individual contributions of *E* and *I*, *E*-only showed stronger predictive power than *I*-only. This aligns with the findings in Section 4.2 that Expansion is connected to age through the length-mediated pathway. However, combining both metrics yielded the best performance, confirming they provide complementary developmental signals that generalize beyond explanation to prediction.

When we evaluate simple age prediction with LLMs using only children's utterances, `Mixtral` and `Gemini` attain MAEs of approximately 1.3 and 1.03, respectively, predicting age in decimal years

(conversion details in Appendix C). Predicting age down to the month level is a challenging task requiring fine-grained discrimination. Given these low errors, we cannot rule out contamination from potential pretraining exposure to the CHILDES dataset (we present supporting checks in Appendix F). Therefore, because the low errors on CHILDES may reflect prior exposure during pretraining, we cannot ascertain the models' true ability to predict age, and it remains uncertain whether they would generalize beyond CHILDES. In contrast, our approach leverages general developmental concepts rather than memorizing dataset-specific patterns, and thus should be more robust.

4.4 Semantic Sensitivity

Section 4.2 decomposed age effects and showed that Expansion is length-mediated while Independence reflects contextual dependence. This section examines whether our metrics are semantically sensitive when length is not a factor. Specifically, we ask the following question: Within utterances matched on length and *PT*, do Expansion and Independence vary with discourse structure (e.g., causality, contrast, initiative) in ways that a length baseline such as MLU does not? By isolating semantic/discourse markers under fixed length and *PT*, we test our metrics' ability to capture meaning-level differences beyond length, complementing the developmental validity in Section 4.2.

Baseline (LMM)	Model	Effect Type	Causal	Contrast	Initiative
MLU	S1	β	$\approx 0\ (P > .10)$	$\approx 0\ (P > .10)$	$\approx 0\ (P > .10)$
	S2		**3.03** $(P < .001)$	**3.42** $(P < .001)$	**1.98** $(P < .001)$
Our metrics (CLMM)					
Expansion (E)	S1	OR	**12.10** $(P < .001)$	**6.73** $(P < .001)$	**3.44** $(P < .001)$
	S2		**4.49** $(P < .001)$	**2.36** $(P < .001)$	**1.86** $(P < .001)$
Independence (I)	S1	OR	**2.54** $(P < .001)$	**1.77** $(P < .001)$	**3.85** $(P < .001)$
	S2		**2.11** $(P < .001)$	**1.84** $(P < .001)$	**4.00** $(P < .001)$

Table 4: Comparison of baselines and our metrics across discourse marker types. All analyses use *PT* and *E*/*I* scores annotated by Gemini 2.5 Pro. Effects that are significant are bolded ($P < .001$); non-significant effects are shown as $P > .10$.

Setup In Equation 3 (S1), we hold length and *PT* fixed to isolate and test the pure semantic effect of the presence of discourse markers M_k on the scores. This design allows us to attribute any remaining differences purely to the presence or absence of markers (see Appendix Table 13). We discretize utterance length (in words) into three bands, *LenBin* $\in \{1\text{-}3, 4\text{-}6, 7\text{-}10\}$. In Equation 4 (S2), we deliberately relax the cell-wise constraint to test robustness under mixed lengths. Equations 3, 4 define the shared linear predictor η. [3]

$$Y = \beta_0 + \sum_k \beta_{M_k} M_k + u_{\text{child}} + u_{PT:LenBin} + \varepsilon \quad (3)$$

$$Y = \beta_0 + \sum_k \beta_{M_k} M_k + \beta_w L_z^{(w)} + \beta_b L_z^{(b)} \\ + \boldsymbol{\beta}_{PT}^\top PT + u_{\text{child}} + \varepsilon \quad (4)$$

If odds ratios that are large in S1 remain clearly > 1 and significant in S2, the signal is not an artifact of cell conditioning but genuinely survives under realistic length variation. Some attenuation is expected because length heterogeneity is reintroduced. We exclude vocd-D, FKGL, and GFI, as these measures assume concatenation or sufficient text for stable estimation.

Analysis For E in Table 4, causal markers yield an OR of approximately 12.10, meaning that within the same length–*PT* cell, the presence of causal connectives increases the odds of moving to a higher category by roughly twelvefold. In S2, this attenuates to 4.49, but the direction and significance remain, indicating that causal linking boosts perceived E even after reintroducing length heterogeneity. For I, causal markers yield ORs of 2.54 in S1 and 2.11 in S2, both clearly > 1, suggesting that causality reflects more than verbosity. It aligns

with the child's initiative in advancing discourse. Similarly, contrast markers raise both E and I at equal length in S1 ($E \approx 6.7\times, I \approx 1.8\times$), and their effects remain substantial in S2. Initiative markers are positive for E as well, but their impact is strongest on I ($\approx 4.0\times$) and remains robust in S2, consistent with the idea that proposals or tag questions strongly signal discourse advancement. Turning to MLU, in S1, it does not respond to marker presence at all, as expected when length is fixed within cells. In S2, positive coefficients emerge, but this reflects a side effect of length increase rather than sensitivity to meaning, as utterances with such markers tend to be longer when length is not constrained. In other words, MLU responds to utterance length but not to semantic content or discourse strategy, functioning as a length-dependent proxy rather than a measure of meaning.

4.5 Human Evaluation

We evaluate how well the proposed Expansion and Independence scores align with human judgments. From the adult–child dialogues, we draw a stratified sample of 300 turns and obtain annotations from three independent expert raters using the same rubric as the LLM. To prevent anchoring bias, raters did not have access to the models' predictions. We report human–human inter-annotator agreement (Table 5) to assess the reliability of the rubric using Krippendorff's α. Krippendorff's α values (Table 5) suggest substantial agreement for Expansion and Independence, and moderate agreement for PT. For model–human alignment, we do not treat machine labels as interchangeable with human annotations. Instead, we evaluate the model's ability to reproduce human ratings using an error-sensitive metric appropriate for ordinal targets (average MAE to each rater), following recommendations for ordinal evaluation (Chen et al., 2021)

and prior work cautioning against agreement measures that assume interchangeable annotators for machine-generated labels (Elangovan et al., 2025). This yields MAE = 0.95/1.12 (Expansion/Independence) for Gemini 2.5 Pro and 1.16/1.45 for Mixtral-8x7B-Instruct.

Metric	Krippendorff's α	Fleiss' κ	Avg Cohen's κ
PT	0.602 [0.556, 0.647]	0.601 [0.555, 0.646]	0.601 [0.556, 0.647]
Expansion	0.858 [0.827, 0.880]	–	–
Independence	0.759 [0.701, 0.809]	–	–

Table 5: Human–human inter-annotator agreement among three expert raters. We report Krippendorff's α for ordinal tasks (Expansion/Independence) and α/κ for nominal PT. Brackets indicate 95% confidence intervals.

5 Broader Impacts

Pedagogical Evaluation and Tutoring Optimization Our metrics provide pedagogically meaningful signals for evaluating and optimizing educational dialogue systems. Unlike standard overlap-based metrics such as BLEU and ROUGE, which encourage spoon-feeding behaviors, our framework evaluates interactional outcomes by measuring whether a system elicits more Expanded and Independent responses from learners. In education, the goal is not merely to provide correct answers, but to ask questions that encourage learners to think. For example, scaffolding questions that prompt reasoning are often pedagogically preferable to direct answers, yet are penalized by standard metrics for low lexical overlap with reference responses. In contrast, our approach focuses on interactional outcomes rather than surface form. As shown in Table 1, the proposed metrics capture systematic differences between questioning strategies (e.g., open vs. closed prompts). For instance, a tutoring agent could be rewarded for using Elaboration-type prompts that elicit richer child contributions (median Expansion = 4), rather than relying on constrained formats such as Yes/No or Display questions (median Expansion = 2, Independence IQR = 0), which limit opportunities for autonomous child talk. Moreover, they can serve as dense, turn-level reward signals for reinforcement learning, enabling tutoring agents to prioritize scaffolding behaviors such as guided questioning rather than overly verbose instruction.

Auditing Synthetic Dialogues and AI Safety With rising interest in simulating children (Mil-

ička et al., 2024), recent work has also focused on constructing and analyzing synthetic child–adult dialogues (Feng et al., 2024). However, existing studies primarily focus on task correctness and do not verify whether generated conversations reflect developmentally natural discourse. This raises a critical question of who evaluates whether synthetic data captures authentic child discourse characteristics. We argue that our metrics provide a promising approach for this essential quality control and auditing role. By verifying whether synthetic dialogues adhere to age-appropriate Expansion and Independence distributions, researchers can filter unrealistic mimicry and ensure the quality of training data.

Beyond data auditing, our results also have implications for AI safety and age assurance. AI providers (e.g., OpenAI's age prediction) actively seek robust age assurance methods to enforce safety policies. We demonstrate a strong correlation between discourse-level metrics and age (Section 4.3), suggesting that these interpretable linguistic features could complement keyword-based filters by enabling platforms to distinguish child users based on language maturity rather than surface lexical cues.

6 Conclusion

We introduced two context-aware metrics, **Expansion** and **Independence**, that assess how a child's utterance organizes meaning and advances discourse, not just how long it is. Conditioning on the **Previous Adult Utterance Type**, *E/I* show clear distributions, remain developmentally valid, capture semantic/discourse structure, and improve age prediction over baselines. Our metrics refocus evaluation on context-sensitive discourse and enable multimodal and cross-linguistic extensions.

Limitations

We analyze orthographic transcripts only, so prosody and acoustic cues are not captured and are left for future work. Our results come from North American English corpora and may not generalize to other languages. For the LLM baselines when predicting age, the models may have encountered highly similar conversations during pretraining.

Ethical Considerations

We use publicly available, de-identified child–adult transcripts (CHILDES/TalkBank) under their terms

of use and make no attempt to re-identify individuals. Age-related analyses are used to validate the developmental sensitivity of our metrics, not to infer or verify the identity of individual users. The metrics are not intended for diagnostic, clinical, or high-stakes decisions about children. LLM-as-a-judge scoring may encode societal or dataset biases. To address this, we release prompts/code, report rubric details, and encourage independent replication and auditing. We avoid releasing any sensitive raw text beyond what licenses permit, and we caution against using these metrics to gate access or impose labels on children in real-world systems.

Acknowledgments

References

Kirsten Abbot-Smith, Julie Dockrell, Alexandra Sturrock, Danielle Matthews, and Charlotte Wilson. 2023. Topic maintenance in social conversation: What children need to learn and evidence this can be taught. *First Language*, 43(6):614–642.

Ahmed Alduais, Issa Al-Qaderi, and Hind Alfadda. 2022. Pragmatic language development: Analysis of mapping knowledge domains on how infants and children become pragmatically competent. *Children*, 9(9).

Thomas Beelen, Roeland Ordelman, Khiet P. Truong, Vanessa Evers, and Theo Huibers. 2024. A conversational robot for children's access to a cultural heritage multimedia archive. In *Advances in Information Retrieval*, pages 144–151, Cham. Springer Nature Switzerland.

Lois Bloom, Lorraine Rocissano, and Lois Hood. 1976. Adult-child discourse: Developmental interaction between information processing and linguistic knowledge. *Cognitive Psychology*, 8(4):521–552.

Martin D. S. Braine. 1963. The ontogeny of english phrase structure: The first phase. *Language*, 39(1):1–13.

Roger Brown. 1973. Development of the first language in the human species. *American psychologist*, 28(2):97.

Maureen A. Callanan, Cristine H. Legare, David M. Sobel, Garrett J. Jaeger, Susan Letourneau, Sam R. McHugh, Aiyana Willard, Aurora Brinkman, Zoe Finiasz, Erika Rubio, Adrienne Barnett, Robin Gose, Jennifer L. Martin, Robin Meisner, and Janella Watson. 2020. Exploration, explanation, and parent–child interaction in museums. *Monographs of the Society for Research in Child Development*, 85(1):7–137.

David Chen, Maury Courtland, Adam Faulkner, and Aysu Ezen-Can. 2021. Error-sensitive evaluation for ordinal target variables. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 189–199, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael M Chouinard. 2007. Children's questions: a mechanism for cognitive development. *Monographs of the Society for Research in Child Development*, 72 1:vii–ix, 1–112; discussion 113–26.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Fredrik deBoer. 2014. Evaluating the comparability of two measures of lexical diversity. *System*, 47:139–145.

Jinming Du and Ben Kei Daniel. 2024. Transforming language education: A systematic review of ai-powered chatbots for english as a foreign language speaking practice. *Computers and Education: Artificial Intelligence*, 6:100230.

Sarita L. Eisenberg, Tara McGovern Fersko, and Cheryl Lundgren. 2001. The use of mlu for identifying language impairment in preschool children. *American Journal of Speech-Language Pathology*, 10(4):323–342.

Aparna Elangovan, Lei Xu, Jongwoo Ko, Mahsa Elyasi, Ling Liu, Sravan Bodapati, and Dan Roth. 2025. Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation and llm-as-a-judge. *Preprint*, arXiv:2410.03775.

Jacqueline Evers-Vermeul and Ted Sanders. 2011. Discovering domains – on the acquisition of causal connectives. *Journal of Pragmatics*, 43(6):1645–1662. Postcolonial pragmatics.

Maria-José Ezeizabarrena and Iñaki García Fernández. 2018. Length of utterance, in morphemes or in words?: Mlu3-w, a reliable measure of language development in early basque. *Frontiers in Psychology*, 8.

Steven Y. Feng, Noah D. Goodman, and Michael C. Frank. 2024. Is child-directed speech effective training data for language models? *Preprint*, arXiv:2408.03617.

Robyn Fivush, Catherine A Haden, and Elaine Reese. 2006. Elaborating on elaborations: Role of maternal reminiscing style in cognitive and socioemotional development. *Child Development*, 77(6):1568–1588.

Virginia C. Mueller Gathercole and Kemp Williams. 1994. A. radford, syntactic theory and the acquisition of english syntax: the nature of early child grammars of english. oxford: Basil blackwell, 1990. pp. vii + 311.1. *Journal of Child Language*, 21(2):489–500.

Jerry Greenfield and Danielle McNamara. 2008. Assessing text readability using cognitively based indices. *TESOL Quarterly*, 42.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

R. Gunning. 1952. *The Technique of Clear Writing*. McGraw-Hill.

Ling-Yu Guo and Sarita Eisenberg. 2015. Sample length affects the reliability of language sample measures in 3-year-olds: Evidence from parent-elicited conversational samples. *Language, Speech, and Hearing Services in Schools*, 46(2):141–153.

V. Heather Fritzley, Rod C. L. Lindsay, and Kang Lee. 2013. Young children's response tendencies toward yes–no questions concerning actions. *Child Development*, 84(2):711–725.

Carla W. Hess, Holly T. Haug, and Richard G. Landry. 1989. The reliability of type-token ratios for the oral language of school age children. *Journal of Speech, Language, and Hearing Research*, 32(3):536–540.

Carla W. Hess, Karen M. Sefton, and Richard G. Landry. 1986. Sample size and type-token ratios for oral language of preschool children. *Journal of Speech, Language, and Hearing Research*, 29(1):129–134.

Jeffrey Kallay and Melissa Redford. 2020. Clause-initial and usage in a cross-sectional and longitudinal corpus of school-age children's narratives. *Journal of Child Language*, 48:1–22.

Linda Kelly, Elizabeth Nixon, and Jean Quigley. 2022. It's your turn: The dynamics of conversational turn-taking in father-child and mother-child interaction. *Language Development Research*, 2:37–68.

Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Lasha Labadze, Maya Grigolia, and Lela Machaidze. 2023. Role of ai chatbots in education: systematic literature review. *International Journal of Educational Technology in Higher Education*, 20.

William Labov and Joshua Waletzky. 1997. Narrative analysis: Oral versions of personal experience. *Journal of Narrative and Life History*, 7(1-4):3–38.

Kathryn A. Leech, Virginia C. Salo, Meredith L. Rowe, and Natasha J. Cabrera. 2013. Father input and child vocabulary development: The importance of wh questions and clarification requests. *Seminars in Speech and Language*, 34(4):249–259.

Chaohao Lin, Ou Bai, Jennifer Piscitello, Emily L. Robertson, Brittany Merrill, Kellina Lupas, and William E. Pelham. 2023. Assessment of parent–child interaction quality from dyadic dialogue. *Applied Sciences*, 13(20).

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

Retief Louw, Emma Sharratt, Febe de Wet, Christiaan Jacobs, Annelien Smith, and Herman Kamper. 2025. Automatically assessing oral narratives of afrikaans and isixhosa children. *Preprint*, arXiv:2507.13205.

Rufan Luo, Lillian R. Masek, Rebecca M. Alper, and Kathy Hirsh-Pasek. 2022. Maternal question use and child language outcomes: The moderating role of children's vocabulary skills and socioeconomic status. *Early Childhood Research Quarterly*, 59:109–120.

Brian Macwhinney. 2000. The childes project: Tools for analyzing talk (third edition): Volume i: Transcription format and programs, volume ii: The database. *Computational Linguistics*, 26:657–657.

Brian MacWhinney. 2014. *The CHILDES Project: Tools for Analyzing Talk, Volume I: Transcription Format and Programs*. Psychology Press, New York.

Lillian R. Masek, Brianna T.M. McMillan, Sarah J. Paterson, Catherine S. Tamis-LeMonda, Roberta Michnick Golinkoff, and Kathy Hirsh-Pasek. 2021. Where language meets attention: How contingent interactions promote learning. *Developmental Review*, 60:100961.

Danielle Matthews, Hannah Biney, and Kirsten Abbot-Smith. 2018. Individual differences in children's pragmatic ability: A review of associations with formal language, social cognition, and executive functions. *Language Learning and Development*, 14.

Allyssa McCabe. 1996. Relating events in narrative: a crosslinguistic developmental study. *Journal of Child Language*, 23:715 – 723.

Philip M. McCarthy and Scott Jarvis. 2007. vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4):459–488.

G McKee, D Malvern, and B Richards. 2000. Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing*, 15(3):323–338.

Jiří Milička, Anna Marklová, Klára VanSlambrouck, Eva Pospíšilová, Jana Šimsová, Samuel Harvan, and Ondrej Drobil. 2024. Large language models are able to downplay their cognitive abilities to fit the persona they simulate. *PLOS ONE*, 19.

J.F. Miller, K. Andriacchi, and A. Nockerts. 2012. *Assessing Language Production Using SALT Software: A Clinician's Guide to Language Sample Analysis*. SALT Software.

Kimberly A. Murphy, Alisha P. Springle, Mollee J. Sultani, and Autumn McIlraith. 2022. Predicting language performance from narrative language samples. *Journal of Speech, Language, and Hearing Research*, 65(2):775–784.

Anat Ninio and Catherine E. Snow. 1996. *Pragmatic Development*, 1 edition. Routledge, New York.

Marilyn A. Nippold, Linda J. Hesketh, Jill K. Duthie, and Tracy C. Mansfield. 2005. Conversational versus expository discourse. *Journal of Speech, Language, and Hearing Research*, 48(5):1048–1064.

David Pagmar, Kirsten Abbot-Smith, and Danielle Matthews. 2022. Predictors of children's conversational contingency. *Language Development Research*, 2:139–179.

Raising Children Network. 2025. Language development in children: what you need to know (0–8 years). https://raisingchildren.net.au/babies/development/language-development/language-development-0-8. Accessed: 2025-10-05.

Nan Bernstein Ratner, Youngjin Han, and Ji Seung Yang. 2024. Should we stop using lexical diversity measures in children's language sample analysis? *American Journal of Speech-Language Pathology*, 33(4):1986–2001.

Janice Redish. 2000. Readability formulas have even more limitations than klare discusses. *ACM J. Comput. Doc.*, 24(3):132–137.

Mabel L. Rice, Filip Smolik, Denise Perpich, Travis Thompson, Nathan Rytting, and Megan Blossom. 2010. Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments. *Journal of Speech, Language, and Hearing Research*, 53(2):333–349.

Brian Richards. 1987. Type/token ratios: what do they really tell us? *Journal of child language*, 14:201–9.

Meredith L. Rowe. 2012. A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child development*, 83 5:1762–74.

Meredith L. Rowe, Kathryn A. Leech, and Natasha Cabrera. 2017. Going beyond input quantity: Wh-questions matter for toddlers' language and cognitive development. *Cognitive Science*, 41(S1):162–179.

Meredith L. Rowe and Catherine E. Snow. 2019. Analyzing input quality along three dimensions: interactive, linguistic, and conceptual. *Journal of Child Language*, 47:5 – 21.

Sherry Ruan, Angelica Willis, Qianyao Xu, Glenn M. Davis, Liwei Jiang, Emma Brunskill, and James A. Landay. 2019. Bookbuddy: Turning digital materials into interactive foreign language lessons through a voice chatbot. In *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale*, L@S '19, New York, NY, USA. Association for Computing Machinery.

Dorothé Salomo, Elena V. M. Lieven, and Michael Tomasello. 2013. Children's ability to answer different types of questions. *Journal of child language*, 40 2:469–91.

Emma Sharratt, Annelien Smith, Retief Louw, Daleen Klop, Febe de Wet, and Herman Kamper. 2025. Feature-based analysis of oral narratives from afrikaans and isixhosa children. *Preprint*, arXiv:2507.13164.

Catherine Snow. 1977. Mothers' speech research: From input to interaction.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

M.C. Templin. 1975. *Certain Language Skills in Children: Their Development and Interrelationships*. Certain Language Skills in Children. Bloomsbury Academic.

Jill C. Thorson, Jill M. Trumbell, and Kimberly Nesbitt. 2024. Caregiver and child question types during a museum interaction. *Frontiers in Psychology*, Volume 15 - 2024.

Anne Thwaite. 2019. Halliday's view of child language learning: Has it been misinterpreted? *Australian Journal of Teacher Education*, 44:42–56.

Claudio O. Toppelberg and Brian Collins. 2004. Constructing a language: A usage-based theory of language acquisition. *Journal of the American Academy of Child & Adolescent Psychiatry*, 43(10):1305–1306.

Trang Tran, Morgan Tinkler, Gary Yeung, Abeer Alwan, and Mari Ostendorf. 2020. Analysis of disfluency in children's speech. *CoRR*, abs/2010.04293.

Jeanine Treffers-Daller, Patrick Parslow, and Shirley Williams. 2016. Back to basics: How measures of lexical diversity can help discriminate between cefr levels. *Applied Linguistics*, 39(3):302–327.

Kristin Valentino, Leah C. Hibel, E. Mark Cummings, Amy K. Nuttall, Michelle Comas, and Christina G. McDonnell. 2015. Maternal elaborative reminiscing mediates the effect of child maltreatment on behavioral and physiological functioning. *Development and Psychopathology*, 27(4pt2):1515–1526.

Mark VanDam, Lauren Thompson, Elizabeth Wilson-Fowler, Sarah Campanella, Kiley Wolfenstein, and Paul De Palma. 2022. Conversation initiation of mothers, fathers, and toddlers in their natural home environment. *Computer Speech & Language*, 73:101338.

R. Veen. 2011. The acquisition of causal connectives: the role of parental input and cognitive complexity. *Sedimentology*.

Brenda M. Wright. 2016. Display and referential questions: Effects on student responses. *Nordic Journal of English Studies*, 15(4):160–189.

Ji Seung Yang, Carly Rosvold, and Nan Bernstein Ratner. 2022. Measurement of lexical diversity in children's spoken language: Computational and conceptual considerations. *Frontiers in Psychology*, Volume 13 - 2022.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Mohammad Zohrabi, Massoud Yaghoubi-Notash, and Yousefzadeh Sanaz. 2014. Teachers' use of display vs. referential questions across different proficiency levels. *International Journal of Applied Linguistics & English Literature*, 3.

A Previous Adult Utterance Type Specification and Related Research

Supporting Literature We define Previous Adult Utterance Type (*PT*) as a taxonomy of adult utterances, covering both *question* and *non-question* categories based on the adult's communicative intent (e.g., eliciting information vs. providing feedback). Our taxonomy draws on constructs from prior work in parent–child interaction and dialogue-act research, where adult question types and scaffolding strategies have been shown

to influence children's responses and language development. Table 6 lists our categories with representative citations.

Prompt To support reproducibility, we provide the full system prompt used for LLM-based classification of adult turns (Table 11). The prompt assigns a single, mutually exclusive label to each adult utterance based on its primary communicative intent. It determines whether an utterance is a question from content and structure rather than punctuation, and applies question and non-question taxonomies grounded in prior work on parent–child interaction and conversational scaffolding. Ambiguous or unintelligible turns receive a special label to prevent spurious classifications.

B Expansion and Independence Specification and Related Research

Supporting Literature We define Independence (*I*) and Expansion (*E*) as complementary dimensions of child language, grounded in prior developmental research. *I* measures conversational agency from imitation to pragmatic leadership, while *E* tracks utterance elaboration from single words to complex narratives. Tables 7 and 8 list the foundational studies underlying each score level.

Prompt We provide the full system prompt used for LLM-based scoring of child utterances on Independence and Expansion (Table 14). The prompt first filters out hesitation-only responses, then scores content-based utterances.

C Dataset Preprocessing

We use 20 corpora from CHILDES: *Bliss, Bohannon, Braunwald, Brown, Champaign, Clark, Demetras1, Demetras2, EllisWeismer, Gleason, HSLLD, Kuczaj, McCune, Nelson, NewmanRatner, OCSC, Post, Sachs, Suppes, Warren*. For each corpus, we extract child and adult utterances from the transcripts, align them by speaker turns, and pass them to the analysis pipeline.

Transcript cleaning and tokenization We apply three preprocessing steps. First, we retain standard CHILDES markup while removing symbols irrelevant to analysis, including tokens consisting only of xxx (marking unintelligible speech). Second, for underscore-compounded words (under_score_word), we replace underscores with spaces to enable space-based tokeniza-

PT	Representative citations
Display Question	Wright (2016)
Yes/No Question	Stolcke et al. (2000); Wright (2016)
Choice Question	Thorson et al. (2024); Callanan et al. (2020)
Referential Question	Rowe et al. (2017); Wright (2016)
Clarification Request	Clark (2020)
Elaboration Prompt	Fivush et al. (2006); Callanan et al. (2020)
Confirmation Check	Clark (2020); Leech et al. (2013)
Topic Introduction	Callanan et al. (2020); Rowe and Snow (2019)
Imperative/Request	Stolcke et al. (2000)
Feedback/Evaluation	Rowe (2012); Lin et al. (2023)
Self-Talk/Thinking Aloud	Stolcke et al. (2000)

Table 6: Previous Adult Utterance Type (*PT*) taxonomy and representative prior work.

tion (e.g., under_score_word → under score word).

Session and utterance filtering We remove segments that provide weak predictive signal or may introduce structural bias:

- **Memorized or recited material**: songs, chants, verse recitation, alphabet or number reading

- **Read-aloud or elicited repetition**: shared book reading, read-aloud tasks, explicit repetition prompts

- **Word-list production**: picture naming, word lists, nonword or syllable repetition

- **Fixed scripts**: play-acting driven by scripted dialogue

We also exclude empty lines, non-utterance lines, and lines containing only noise markers such as xxx. Utterances consisting solely of hesitation or filler sounds (*um, uh, hm*) are treated as *hesitation-only*, assigned Expansion=0, Independence=0, and then excluded from analysis.

Age conversion Age notations in transcripts (P#Y#M, 6;9.12, and related formats) are converted to decimal years. For example, P6Y07M → 6.58 years and 4;7.0 → 4.58 years.

D Discourse Markers

We use a set of surface-level discourse markers to categorize pragmatic and structural cues in adult utterances. These markers guide how child responses are scored for Expansion and Independence by capturing cues such as causal reasoning, contrastive organization, and conversational initiative. Table 13 lists the marker categories and corresponding lexical patterns.

E Age Prediction

LLM Age Prediction Prompt The following system prompt (Table 12) was used for the large language model evaluation described in Section 4.3. The model inferred a child's age in years and months from a transcript containing only the child's utterances.

Training Details All models were trained using 5-fold GroupKFold cross-validation to prevent child-level data leakage. Hyperparameters were selected based on validation MAE. For each fold, we computed mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (R^2), reporting mean values and 95% confidence intervals. For the non-linear model, we used scikit-learn's GradientBoostingRegressor with 500 estimators, a learning rate of 0.05, maximum depth of 3, and a fixed random state.

95% CIs Tables 9 and 10 show 95% confidence intervals (CIs).

F Potential Pretraining Exposure

We observed that both Mixtral-8x7B and Gemini 2.5 Pro achieve notably low mean absolute errors (MAE ≈ 1.3 and 1.03 years, respectively) on age prediction over CHILDES transcripts. While such precision may indicate genuine linguistic generalization, we cannot rule out the possibility that it is partially influenced by prior exposure to overlapping CHILDES–style transcripts during pretraining. To probe this possibility, we qualitatively examined a small set of responses (Table 15) where models hypothesized specific corpora from contextual cues or produced meta-level statements suggesting familiarity with the CHILDES ecosystem. Importantly, these examples are anecdotal and do not constitute evidence of verbatim memorization,

Expansion score	Representative citations
1	Bloom et al. (1976)
2	Braine (1963); Gathercole and Williams (1994)
3	Gathercole and Williams (1994); Bloom et al. (1976)
4	Nippold et al. (2005)
5	Nippold et al. (2005); Rowe and Snow (2019)
6	Kallay and Redford (2020); Nippold et al. (2005)
7	Evers-Vermeul and Sanders (2011); Valentino et al. (2015)
8	McCabe (1996); Fivush et al. (2006); Labov and Waletzky (1997)
9	Nippold et al. (2005); Evers-Vermeul and Sanders (2011)
10	McCabe (1996); Fivush et al. (2006); Nippold et al. (2005)

Table 7: Mapping of E scores (1–10) to representative prior work (no MLU citations).

Independence score	Representative citations
1	Bloom et al. (1976)
2	Bloom et al. (1976)
3	Clark (2020); Heather Fritzley et al. (2013)
4	Wright (2016); Salomo et al. (2013); Heather Fritzley et al. (2013)
5	Clark (2020); Leech et al. (2013)
6	Chouinard (2007); Rowe et al. (2017)
7	Fivush et al. (2006); Rowe and Snow (2019); Valentino et al. (2015)
8	McCabe (1996); Fivush et al. (2006)
9	Abbot-Smith et al. (2023); Clark (2020); VanDam et al. (2022)
10	Masek et al. (2021); Abbot-Smith et al. (2023)

Table 8: Mapping of I scores (1–10) to representative prior work (no MLU citations).

and self-reported exposure is not independently verifiable. Because pretraining data for these models are not fully disclosed, we interpret the results as consistent with potential dataset exposure rather than confirmed contamination.

Type	Feature set	Linear Regression			Gradient Boosting		
		MAE	MSE	R^2	MAE	MSE	R^2
Baselines	MLU	1.53 [1.39–1.68]	3.84	0.17 [0.12–0.21]	1.52 [1.34–1.70]	4.25	0.07 [-0.28–0.38]
	MLU + *PT*	1.51 [1.36–1.66]	3.78	0.18 [0.14–0.22]	1.50 [1.32–1.69]	4.15	0.09 [-0.25–0.38]
	vocd-D	1.76 [1.65–1.87]	4.58	0.01 [0.00–0.02]	1.81 [1.73–1.90]	4.87	-0.06 [-0.12–0.00]
	vocd-D + *PT*	1.72 [1.61–1.84]	4.47	0.03 [0.02–0.04]	1.76 [1.66–1.86]	4.63	0.00 [-0.05–0.03]
	FKGL	1.76 [1.61–1.84]	4.55	0.02 [0.02–0.04]	1.73 [1.57–1.87]	4.56	0.01 [-0.03–0.07]
	FKGL + *PT*	1.73 [1.61–1.84]	4.46	0.03 [0.02–0.04]	1.71 [1.57–1.85]	4.47	0.03 [-0.03–0.07]
	GFI	1.66 [1.54–1.78]	4.22	0.09 [0.07–0.11]	1.52 [1.39–1.64]	3.80	0.18 [0.15–0.22]
	GFI + *PT*	1.64 [1.52–1.76]	4.17	0.10 [0.08–0.12]	1.52 [1.39–1.64]	3.75	0.19 [0.15–0.22]
Ours	*E + I*	1.20 [1.11–1.27]	2.75	0.40 [0.36–0.46]	1.23 [1.12–1.31]	2.97	0.36 [0.29–0.45]
	E + I + PT	**1.19**[*] [1.12–1.28]	**2.71**[*]	**0.41**[*] [0.35–0.46]	**1.22**[*] [1.12–1.31]	**2.89**[*]	**0.37**[*] [0.29–0.45]
	E	1.20 [1.12–1.28]	2.75	0.40 [0.35–0.46]	1.23 [1.09–1.38]	2.97	0.36 [0.29–0.42]
	E + PT	1.20 [1.11–1.28]	2.73	0.41 [0.35–0.46]	1.23 [1.09–1.37]	2.97	0.36 [0.29–0.41]
	I	1.34 [1.25–1.43]	3.17	0.31 [0.27–0.36]	1.35 [1.23–1.48]	3.38	0.26 [0.15–0.36]
	I + PT	1.33 [1.24–1.42]	3.13	0.32 [0.28–0.36]	1.34 [1.23–1.47]	3.33	0.27 [0.16–0.37]

Table 9: Mixtral-8x7B-Instruct results. 95% confidence intervals (CIs, shown in blue) are reported.

Type	Feature set	Linear Regression			Gradient Boosting		
		MAE	MSE	R^2	MAE	MSE	R^2
Baselines	MLU	1.42 [1.36–1.48]	3.47	0.23 [0.15–0.31]	1.55 [1.44–1.66]	4.10	0.09 [0.08–0.22]
	MLU + *PT*	1.40 [1.35–1.46]	3.41	0.25 [0.17–0.32]	1.50 [1.41–1.55]	3.91	0.14 [0.11–0.25]
	vocd-D	1.75 [1.67–1.83]	4.55	0.00 [-0.01–0.00]	1.78 [1.73–1.82]	4.66	-0.03 [-0.06–0.02]
	vocd-D + *PT*	1.71 [1.63–1.79]	4.41	0.03 [0.02–0.03]	1.70 [1.64–1.74]	4.35	0.04 [0.00–0.09]
	FKGL	1.72 [1.64–1.79]	4.40	0.03 [0.02–0.04]	1.67 [1.59–1.75]	4.27	0.06 [0.01–0.10]
	FKGL + *PT*	1.69 [1.61–1.76]	4.31	0.05 [0.04–0.06]	1.63 [1.56–1.72]	4.13	0.09 [0.03–0.12]
	GFI	1.57 [1.49–1.65]	3.92	0.14 [0.10–0.17]	1.43 [1.38–1.49]	3.43	0.25 [0.19–0.29]
	GFI + *PT*	1.56 [1.48–1.64]	3.87	0.15 [0.11–0.18]	1.40 [1.34–1.46]	3.31	0.27 [0.22–0.31]
Ours	*E + I*	1.14 [1.11–1.17]	2.41	0.47 [0.43–0.51]	1.05 [0.99–1.06]	2.09	0.54 [0.49–0.61]
	E + I + PT	**1.13**[*] [1.10–1.16]	**2.38**[*]	**0.48**[*] [0.44–0.51]	**1.04**[*] [0.98–1.05]	**2.06**[*]	**0.55**[*] [0.50–0.61]
	E	1.22 [1.18–1.25]	2.76	0.39 [0.36–0.43]	1.14 [1.07–1.23]	2.47	0.46 [0.36–0.51]
	E + PT	1.21 [1.17–1.24]	2.71	0.40 [0.37–0.44]	1.12 [1.09–1.20]	2.42	0.47 [0.37–0.51]
	I	1.24 [1.19–1.28]	2.76	0.39 [0.35–0.42]	1.33 [1.26–1.43]	3.03	0.33 [0.24–0.36]
	I + PT	1.23 [1.18–1.27]	2.72	0.40 [0.36–0.43]	1.33 [1.28–1.41]	3.03	0.33 [0.25–0.36]

Table 10: Gemini 2.5 Pro results. 95% confidence intervals (CIs, shown in blue) are reported.

Each adult utterance is marked as [X-a], where X is a number indicating the turn order. Your task is to classify each of these adult utterances based on the categories below.

Important: You must determine whether an utterance is a question or not based solely on the content and structure, not on punctuation.

First, determine whether the adult's utterance is a **direct question**.

If the utterance is a question, classify it as one of the following Question Types:

- **Display Question**: The adult already knows the answer and expects a specific factual response. (Does not include yes/no or choice questions) e.g., "What color is this apple?"

- **Yes/No Question**: Questions explicitly requiring only a yes or no response. e.g., "Did you go outside today?"

- **Choice Question**: Questions explicitly asking the child to select one option from two or more provided choices. e.g., "Do you want juice or milk?"

- **Referential Question**: Questions genuinely seeking information about the child's personal experiences, opinions, or thoughts, with no predetermined correct answer. e.g., "What did you do at school today?"

- **Clarification Request**: Adult requests the child to repeat or clarify a previous utterance due to misunderstanding or inaudibility. e.g., "What did you say?", "Huh?", "Can you say that again clearly?"

- **Elaboration Prompt**: Questions or statements explicitly prompting the child to expand or provide additional details, reasons, or explanations about something previously mentioned. e.g., "Can you tell me more about that?", "Why did that happen?"

- **Confirmation Check**: Adult explicitly confirms their understanding of the child's previous utterance, usually by rephrasing or summarizing the child's words. e.g., "Did you mean the red one?", "You're saying you went to the park?"

If the utterance is not a question, classify it as one of the following Non-Question Types:

- **Topic Introduction**: Adult introduces a new topic, shares personal experiences, or describes past events to provide context or invite engagement, without asking a question. e.g., "Today I saw a big dog.", "When I was a kid, I visited the mountains every summer."

- **Imperative/Request**: Adult gives a command or request intended to prompt an action, not to obtain information. e.g., "Read please.", "Put that here."

- **Feedback/Evaluation**: Adult provides positive or negative evaluation or feedback on the child's behavior or utterance. e.g., "Good job!", "That's not quite right."

- **Self-Talk/Thinking Aloud**: Adult speaks to themselves or expresses thoughts aloud, not directly addressing the child. e.g., "Hmm, I can't remember where I put that."

Special Cases

- **Ambiguous/Unclear**: When the adult's utterance is completely incomprehensible or unintelligible enough that question type cannot be determined at all. e.g., "xxx", completely garbled speech, or utterances with no meaningful words, such as only "um", "uh", "hm"

Note: "xxx" marks unintelligible speech in the original transcript.

Output Format

Classify each [X-a] adult utterance using the categories above. When multiple types are present, choose the type that represents the main communicative intent of the utterance.

IMPORTANT: Use ONLY the exact categories defined above. Do not create new categories or add explanations in parentheses. Do not include any explanation, commentary, or scoring.

Follow the format exactly as shown below:

Format:

```
    [1-a] - Question Type: Yes/No Question
    [2-a] - Non-Question Type: Topic Introduction
    [3-a] - Question Type: Display Question
    _____

Now, analyze the following conversation: {DIALOGUE}
```

Table 11: System prompt template for classifying Previous Adult Utterance Type.

```
You are given a transcript of only the CHILD's utterances from a conversation. This transcript was
converted from spoken audio. Ignore any empty or whitespace-only lines.

Task:
Estimate the child's age in years and months only.

Output format:
"X years Y months" (no extra words).

Examples:
- 4 years 4 months
- 7 years 0 months

Transcript:
{DIALOGUE}

Output:
```

Table 12: System prompt template used for LLM-based age prediction from child utterances.

Rubric stage	Representative markers / regex patterns
Causal	because, so, so that, since, therefore ...
Contrast	but, however, although, though, whereas, on the other hand, by the way, anyway ...
Initiative	let's, wanna, shall we, how about, maybe we can, do you want to, wanna try ...

Table 13: Discourse marker categories

Each child utterance is labeled with a number in square brackets (e.g., [1], [2], etc.). Please assign Linguistic Independence and Expansion scores to each based on the criteria provided below.

Step 1: Response Content Type

- **Hesitation Only Response**: The child's utterance consists solely of hesitation/filler tokens (e.g., "um," "uh," "hm," "er," "ah") with no propositional or acknowledgment content.
- **Content Response**: The child's utterance contains meaningful words or content.

Step 2: Scoring

- **For Hesitation Only Responses:**
  - Independence: 0 (automatic)
  - Expansion: 0 (automatic)
- **For Content Responses:** Apply the scoring criteria below.

Evaluation Criteria (For Content Responses Only)

Note: When a child's utterance includes both hesitation and real words, exclude the hesitation when evaluating both Linguistic Independence and Expansion scores. e.g., "um I don't know" → Score based only on "I don't know" Focus only on the meaningful content words for all scoring purposes.

---

1) Independence (scoring criteria; 1 = very dependent, 10 = fully independent)

- **Score 1 – Imitation without understanding (complete repetition)**: Directly repeats the adult's words identically; not a response. e.g., Adult: Did you eat? → Child: Did you eat? Key features: verbatim echo, no response content.
- **Score 2 – Partial imitation**: Repeats only the tail of the adult's utterance; does not supply content. e.g., Adult: Did you eat red apple? → Child: Red apple? Key features: echoes final words, no full comprehension.
- **Score 3 – Single acknowledgment only (no answer)**: Acknowledgment tokens only; not an answer to the question. e.g., Yes, No, Okay, Uh-huh Key features: acknowledgment only, no meaningful information.
- **Score 4 – Direct answer only (no initiative)**: Accurately answers what was asked, adds nothing else (no follow-up/evaluation/new info). e.g., Red., At school., Two. Key features: answer present, no initiative.
- **Score 5 – Answer + minimal initiative**: After answering, adds minimal non-exploratory move: closed follow-up, confirmation, clarification, or prompted reason (only after adult asks why). e.g., Red. Which one?, At school—Is that okay?, What?, Because I like it. (after prompt) Key features: minimal initiative, not open-ended, prompted reason allowed.
- **Score 6 – Self-initiated open-ended inquiry**: Independently advances with open-ended exploration (why/how/what-if/compare), with or without an answer. e.g., Red—why is it red?, How does it work?, What if we pick blue? Key features: self-driven inquiry, open-ended.
- **Score 7 – Spontaneous explanation (unprompted justification)**: Gives reasons/justifications without being asked why. e.g., I picked red because it's faster. Key features: unprompted justification.
- **Score 8 – Previous conversation connection**: Links prior talk/events to the current context with temporal continuity. e.g., The friend I told you about—he's back! Key features: across-turn linking, time continuity.
- **Score 9 – Topic management + listener-oriented regulation (≥ 2 features)**: In one turn performs ≥ 2: topic open/shift, listener check, self-monitor/repair, close/return to prior topic. e.g., Anyway, let's talk about the science project now. Does that make sense? I'll say it more slowly. Key features: two or more control acts, functional regulation.
- **Score 10 – Pragmatic leadership with perspective-taking (integrated)**: Leads (proposes/organizes) and adapts to listener needs/state by integrating ≥ 2: plan/proposal/role, perspective-taking, adaptive delivery. e.g., Let's start with the Lego one; if you're tired, we can stop later. I'll explain slowly so you can follow. Key features: plan plus listener adaptation, interactional leadership.

2) Expansion (scoring criteria; 1 = no expansion, 10 = highly elaborated)

- **Score 1 – Single word**: One noun/verb/adjective only. e.g., Dog, Pretty, Dislike Key features: word count=1.

- **Score 2 – Two words**: Modifier+noun or two separate words. e.g., Cute dog, No that Key features: word count=2.
- **Score 3 – 3–4 words, no sentence**: Word combination without full subject–verb structure. e.g., Red pretty apple, Me like this Key features: 3–4 words, no S–V structure.
- **Score 4 – Basic sentence**: One complete clause; no additional information. e.g., The dog runs, I ate rice Key features: single clause.
- **Score 5 – Sentence + modifiers**: Complete clause with 1–2 descriptive words. e.g., The big dog runs fast Key features: descriptive modifiers present.
- **Score 6 – Simple coordination (no causality)**: Uses and/but to coordinate clauses; no causal/conditional meaning. e.g., I ate rice and played, It was fun but tiring Key features: coordination only, no causality.
- **Score 7 – Single cause–effect**: One explicit causal link (because/so/therefore). e.g., It was raining so I stayed inside Key features: exactly one causal link.
- **Score 8 – Detailed sequence**: Chronological sequence of ≥ 3 actions with ordering markers (then/after/first/next). e.g., I went to the park, played on swings, then ate snacks Key features: 3+ ordered events.
- **Score 9 – Complex logical structure**: Includes subordinating logic (if/when/unless/while/than) and remains coherent. e.g., If it rains like today, I'll use an umbrella; if sunny, I'll walk Key features: subordination present.
- **Score 10 – Multi-layered narrative**: Coherent linkage across multiple timeframes/perspectives or abstract connections. e.g., Yesterday's rain reminded me of last summer when we couldn't go camping, which was good because we stayed home and I learned to bake Key features: multiple time/perspective links.

Note: "xxx" marks unintelligible speech in the original transcript.

Output Format

For each child's utterance, provide scores in the following format. Do not include any explanation, or commentary. Follow the format exactly as shown below:

For Content Responses:

[1] - Independence: X, Expansion: Y
[2] - Independence: X, Expansion: Y

For Hesitation Only:

[1] - Independence: 0, Expansion: 0 [Hesitation Only]
[2] - Independence: 0, Expansion: 0 [Hesitation Only]

Example Outputs:

[1] - Independence: 6, Expansion: 5
[2] - Independence: 0, Expansion: 0 [Hesitation Only]
[3] - Independence: 3, Expansion: 2

Now, analyze the following conversation: {DIALOGUE}

Table 14: System prompt template for scoring child responses in child–adult conversational transcripts, defining Linguistic Independence and Expansion.

Case	Transcript Excerpt and Model Responses
(a) HSLLD corpus	*Child: hey. there's a penny and another penny...*
	Adult: it's a magnet. see.
	Mixtral: "Corpus: HSLLD. Multiple speakers (Child, Mother) and object exploration indicate a lab-based design."
	→ Infers the corpus from speaker composition and object-focused interaction, reflecting context-based reasoning.
	Gemini: "This resembles a structured lab play session with a mother and investigator."
	→ Identifies a lab-like play context and speaker roles, suggesting familiarity with similar transcript structures.
(b) Brown corpus	*Adult: do you want some milk?*
	Child: want milk.
	Adult: you can have it.
	Child: mine.
	Mixtral: "Likely Brown corpus–home setting, short child utterances, and caregiver scaffolding."
	→ Attributes the excerpt to Brown based on a home-style exchange, short child turns, and caregiver scaffolding.
	Gemini: "This looks like a caregiver–child exchange during early development, possibly Brown, but uncertain."
	→ Notes early developmental speech patterns and caregiver–child context while explicitly expressing uncertainty.
(c) Bliss corpus	*Adult: your turn!*
	Child: let's move this piece here.
	Adult: good idea.
	Child: I win!
	Mixtral: "Probably Bliss corpus–play-based discourse with initiative markers ('let's', 'your turn')."
	→ Links the excerpt to a play-based interaction using initiative markers such as "let's" and "your turn."
	Gemini: "This resembles CHILDES play-based dialogue involving cooperative games."
	→ Characterizes the dialogue as play-based and cooperative, aligning it with a common CHILDES-style scenario.
(d) Awareness	*Adult: do I get to play farm with you?*
	Child: yeah. they're firefighters...
	Mixtral: "I cannot confirm the corpus but it resembles CHILDES toy-play transcripts (Brown or Bliss)."
	→ Avoids committing to a specific corpus while pointing to toy-play transcript structure typical of CHILDES.
	Gemini: "Consistent with CHILDES toy-play transcripts, but corpus unknown."
	→ Similarly recognizes CHILDES-like toy-play structure and acknowledges that the corpus cannot be determined.
(e) Self-Disclosure	*Child: no no no. they eat chicken wings. and this they're gonna have little ladybugs.*
	Mixtral: (N/A–model did not mention corpus exposure.)
	Gemini: "As a large language model, I don't 'remember' data, but the contents of the Ellis Weismer corpus, along with many other CHILDES corpora, were part of my pretraining dataset."
	→ Self-reports possible exposure to CHILDES corpora during pretraining (not independently verifiable).

Table 15: Qualitative responses from **Mixtral-8×7B-Instruct** and **Gemini 2.5 Pro** that suggest potential pretraining exposure. Models sometimes hypothesize corpus identity from contextual cues or produce meta-level comments about CHILDES.