Measuring Form and Function in Language Models

Anonymous ACL submission

Abstract

We introduce quantitative metrics from child language research to evaluate language models. Our focus is on the formal syntactic and functional discourse properties of determiners in English, which young children acquire early and accurately. We propose *Contextual Alternative Choice* (CAC), a new prompting method which provides targeted tests for both syntactic and discourse knowledge of language. The method enables direct comparison of language models against children, and more importantly, against independently established benchmarks in empirical research. No current model trained on a comparable amount of data simultaneously meet both benchmarks like human children, but some very large models do. We present our results as methodological and technical contributions, with specific emphasis on the cognitive status of language models.[1]

1 Introduction: LLMs and child language

Central to American structuralist linguistics (Harris, 1951; Chomsky, 1955) is the methodological commitment to *distributionalism*: Grammar can be defined formally from a corpus of text (or phonemes) without reference to semantic content. As articulated in *Syntactic Structure* (Chomsky, 1957, p50), the most important goal of linguistic theory under the distributional approach is a *discovery procedure*: "the theory must provide a practical and mechanical method for actually constructing the grammar, given a corpus of utterances." In light of the remarkable success of LLMs, which are primarily trained on text, a set of questions naturally arises: Do LLMs constitute a discovery procedure? Do they encode linguistic knowledge in ways that are comparable to human language users? To what extent can they be regarded as a computational model of language acquisition by children? The BabyLM project tackles such fundamental problems in the science of language and cognition, which may in turn provide insight on building data-efficient models to serve wider linguistic communities (Choshen et al., 2026).

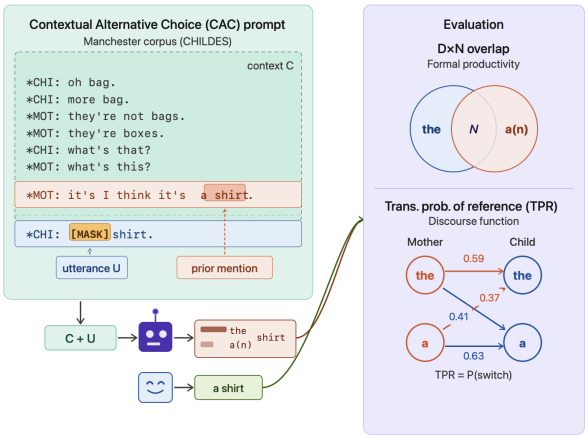

Figure 1: Contextual Alternative Choice (CAC) prompting: A target determiner-noun (D×N) site is identified in caretaker-child dialogue with *the* and *a* as grammatical choices. The model is presented with the full preceding discourse context C and predicts the determiner in target utterance U. The resulting determiner preferences are evaluated against two independently motivated benchmarks: formal productivity (overlap) and discourse-functional appropriateness (Transitional Probability of Reference (TPR) described in Section 2.

To properly approach these questions requires a framework of reference in which child and LLM language learning can be rigorously compared. To do so, we suggest, calls for closer connection with child language research than present. Our starting point is two striking yet underappreciated parallels between language learning by children and LLMs.

First, for very different reasons, the inner mechanisms of human and LLM learning are both poorly understood. There is, quite obviously, very little room for drastic experimentation with human learning. At the same time, LLMs have so far resisted a complete theoretical understanding despite invasive

[1] Our evaluation code and data will be made available on GitHub and linked here once deanonymized.

manipulation of input and model architecture. As such, their external behavior — what children say and what LLMs do — remains the primary source of data for assessing their linguistic capacities. Second, the investigation of children and LLMs both face a data poverty problem. The BabyLM challenge is an exception but frontier LLMs are generally trained with closed-source data. Similarly, while a reasonable amount of child-directed data exists in the public domain (MacWhinney, 2000), it is amassed from many learners, with no comprehensive record for any particular child.

In this paper, we exploit these parallels between child and LLM learning by introducing a new set of investigative tools from recent developments in language acquisition. Our proposal is both methodological and empirical, which we take up in turn.

1.1 Classification vs. generation

To date, the most popular evaluation for LLMs, including the majority in the BabyLM challenge, is based on classification of minimal pairs: grammatical vs. ungrammatical sentences, plausible vs. implausible answers in conceptual understanding and reasoning, correct vs. incorrect pairings between images and captions, etc. Classification datasets, which are easy to scale, are naturally suited for LLMs. However, there are well-known difficulties relating classification tasks to human linguistic behavior and knowledge. For example, the syntactic templates in the widely adopted BLIMP benchmark frequently introduces semantic and pragmatic anomalies (Warstadt et al., 2020). However, the influence of these factors on acceptability judgment does not appear simply additive to syntax, even for binary tasks (Sprouse et al., 2018), making it difficult to relate LLM performance to human behavior (Vázquez Martínez et al., 2023).

The classification task is even more problematic in the cognitive setting of BabyLM: Quite simply, there is no ground truth from child language to compare LLMs against. It has long been familiar to researchers that young children cannot provide reliable metalinguistic judgments (De Villiers and De Villiers, 1974; Thornton, 2021). Indeed, most methods readily applicable to adults cannot extend to children. For example, the celebrated Wug test (Berko, 1958), the most direct test for linguistic productivity (*wug-wugs*), can only reliably elicit correct responses from first graders; the plural rule, however, is already learned by age 2 (Brown, 1973). The gap is due to the heavy metalinguistic demands required for learning and using novel words in the lab.

For these reasons, child language research typically uses naturalistic production data to benchmark language development. For example, the size of children's vocabulary, the simplest component of language, is estimated by only including words that children produce rather than what they (apparently) understand (Fenson et al., 1994). Similarly, the major landmarks of morphological and syntactic acquisition are all established on children's naturalistic production, ranging from the number of morphemes and words per utterance (e.g., mean length of utterance, "one/two word stage") to overgeneralization errors (e.g., *foots*, *goed*, *I died you*) that mark the productivity of rules (Ervin, 1964; Bowerman, 1982). The gold standard for assessment is Brown's 90% criterion (1973), which captures both the formal and functional aspects of language: Acquisition is deem successful if children produce adult-appropriate output in at least 90% of obligatory contexts. Originally established on child English, the criterion has been widely adopted in cross-linguistic research (Slobin, 1985).

1.2 Beyond well-formedness

In recent years, however, Brown's criterion has been under scrutiny. As argued by theorists such as Tomasello (2000), children's adult-like usage may simply result from memorizing, and retrieving, chunks of adult input of varying sizes than a systematic grammar. This should be a familiar refrain in the study of language modeling from n-grams to LLMs: It is difficult to know whether certain linguistic capacity can be attributed to generalizations from, or regurgitation of the training data itself (McCoy et al., 2023; Merrill et al., 2024).

In this paper, we introduce two recent benchmarks for measuring child language designed to overcome the deficiencies in Brown's 90% criterion. Together, they recognize that well-formedness alone may be too weak of a test for linguistic knowledge. These benchmarks, reviewed in Section 2, compare the statistical profile of a corpus, not against the statistical profile of the input, but against the *expected* statistical profile of the target *grammar*. As such, they are applicable to any language sample produced by children and LLMs alike, without needing to access input/training data. Our empirical study focuses on English determiners, a richly researched domain that embodies both formal (syntax) and functional (discourse) proper-

ties of child language. In order make direct comparisons between LLMs and children, we draw inspiration from the familiar cloze task (Taylor, 1953) from psycholinguistics, especially its use at the discourse level (Chihara et al., 1977) and introduce a new LLM prompting method to make targeted grammatical choices in identical contexts as children (Section 3). We use this framework to evaluate determiner use by two-year-olds and also by 45 language models (Section 4) in the BabyLM Project. No model trained on a comparable amount of input data meets both benchmarks although two models trained on very large amount of data do. In Section 5 and 6, we discuss related work and the implication of our work in the broader context of LLM research, with specific emphasis on their status as cognitive models of language.

2 The statistics of determiners

Determiners provide an excellent window into language acquisition. As functional words, determiners' role in language is largely formal, lacking the referential salience and semantic concreteness of content words (nouns, verbs, adjectives, etc.). Nevertheless, determiners are identified as a formal category in infancy and play an important role to establish content word categories such as nouns; see Shi (2023) for review. Determiners also appear early and frequently in child language production and have proven a fertile ground of research.

2.1 Measuring determiner productivity

Children's determiner usage is essentially error-free from the get-go (Valian, 1986): they must have learned an adult-like grammar by Brown's 90% criterion. This conclusion has been challenged in recent years (Tomasello, 2000). Adult-like usage may result from memorization and retrieval of input, which would be lexically specific rather than genuinely productive. For example, the English determiners *the* and *a*[2] are productive because they can interchangeably combine with any singular noun given suitable context. A widely adopted numerical measure, *overlap*, has been used to quantify determiner-noun (henceforth D×N) productivity (Pine and Martindale, 1996). Given a corpus, overlap is the proportion of noun types that appear with both *the* and *a* out of those appearing with either *the* or *a*. The overlap value in child language corpora is usually quite low (20-40%), and considerably lower than their caretakers, leading to claims that children do not have a productive determiner grammar but only a collection of lexically specific collocations from the input.

The overlap measure has generated a large body of quantitative tests for children's grammar but almost all rely on comparisons with, or statistical analysis of, caregiver data to determine the extent of memorization; see Section 5 for details. As such, these methods are inherently limited by the available input data, which at best is only a tiny fraction of the totality. By contrast, a formal statistical test (Yang, 2013) overcomes this obstacle. It derives the expected overlap value generated by a productive grammar, i.e., one that combines D×N independently. The test builds on two key statistical properties of language, one universal and the other specific to D×N in English. First, the frequencies of nouns are assumed to follow Zipf (1949)'s Law. As such, if a corpus of S D×N combinations involves N unique nouns, the expected frequency of the noun of rank r is $S/(rH_N)$ where $H_N = \sum_{i=1}^{N} 1/i$. Second, in D×N combinations, nouns tend to have a "favorite" (more frequent) determiner. For example, *bathroom* greatly favors *the* over *a* but for *bath*, the reverse is true. This imbalance, referred to as *bias* (b), is quantified in the aggregate as follows:

$$b = \frac{\sum_{n \in N} \max(C_{\text{the}\times n},\ C_{\text{a}\times n})}{\sum_{n \in N} (C_{\text{the}\times n} + C_{\text{a}\times n})} \quad (1)$$

where $C_{\text{the/a}\times n}$ is the frequency of *the*/*a* combined with n in the corpus. Children do not need to learn these frequencies: they presumably do not track personal hygiene (*a bath*) or bodily functions (*the bathroom*). Rather, the bias value reflects the vagaries of life in language use, and is thus remarkably stable in the aggregate: 0.82 in the billion-word COCA corpus (Davies, 2008), almost identical to the values in the child and caretaker speech analyzed in Section 4.

These two statistical properties greatly enhance the applicability of the productivity test. For a corpus, one only S, N, and b are need to calculate the expected value of overlap; see Appendix A for details. If a comparison between expected and empirical overlap values from a corpus does not show a significant difference, one can conclude that the corpus is consistent with a productive grammar.

[2] Following the practice in child language, the phonological variant *an* is merged with *a* in the present study.

2.2 Measuring determiner reference

As emphasized in Section 1, linguistic knowledge is more than well-formedness in classification tests. To fully acquire determiners, it is not sufficient to know they freely combine with nouns: their communicative functions are equally important. The determiners *a* and *the* differ in their referential properties (definiteness, specificity, etc.) in context (Lyons, 1999); these functional roles have also been studied in language acquisition. Three-year-olds can interpret determiner reference in context with high accuracy (85%; Maratsos 1976). Analyzing children's naturalistic production, Rozendaal and Baker (2008) conclude that two-year-olds already show adult-like determiner usage with respect to specificity, definiteness and givenness. However, these authors also note that interpreting children's determiner usage requires a reconstruction of the child's mental model of discourse, which is both subjective and laborious.

Gleitman and Yang (2022) propose a statistical measure of determiner function that eliminates interpretative guesswork. They focus on contexts where one speaker uses a determiner in response to a determiner used by their interlocutor when discussing the same topic noun: In general, only one of two options is acceptable. The following two snippets of conversation are instances where the child follows the caretaker with the same determiner; changing it would be highly unnatural albeit formally grammatical.

Mother: Is it **the dog** or the little boy?

Child: **The dog** won't stand up properly.

Mother: I'll make you **a gate**.

Child: Found one. Found **a gate**.

By contrast, the following are instances of determiner change: retaining the same determiner is infelicitous.

Mother: Have you ever had **an itch**?

Child: That's **the itch** over there.

Mother: What happens in **the carwash**, John?

Child: Train's had **a carwash**.

Gleitman and Yang (2022) introduce a statistical measure of determiner discourse function: The *transitional probability of reference* (TPR) is the probability with which a speaker *changes* determiner following interlocutor while holding the topic noun constant, which can be represented as state transitions in Fig 1. For example, in the four transitions above, TPR is 0.5. Children's TPR when talking to caretakers can be compared against adults' TPR when talking to each other, which serves a baseline benchmark. A baseline is motivated by the fact that the frequency of communicative needs (e.g., retaining or shifting reference for definiteness, specificity,e tc.) is largely invariant (Givón, 1983) even when the grammar expressing these functions may change (Komen et al., 2014). Once again, the absence of statistical difference between child and adult TPRs indicates adult-like command of determiners with respect to discourse functions.

3 Methodology

Our main methodological contribution is a prompting scheme that turns child-caretaker dialogue into controlled tests of linguistic knowledge in LLMs. For each target utterance U produced by children, we construct a discourse context C consisting of the preceding dialogue in the recording session, truncated only by the model's context window. Within U, we identify the target position of a grammatical choice and elicit determiner preferences from the model — *the* or *a* — while preserving all surrounding contexts including the noun that the determiner combines with: Hence *Contextual Alternative Choice* (CAC) prompting (see Fig 1).

Our empirical study is based on the Manchester Corpus (Theakston et al., 2001) which contains regular recording sessions of 12 mother-child dyads aged 2-3, when determiners begin to be systematically used in speech. Extrapolating from estimates of child-directed input (Gilkerson et al., 2017), these children will have heard approximately 10-20 million words in total. It is however more likely that the effective input is only 5-10 million words as the first year is primarily phonological development. These quantities provide useful checkpoints for LLMs as models of child acquisition.

The basic workflow of our study is illustrated in Fig 1. Both caretaker and child D×N pairs are extracted from the Manchester Corpus. The caretaker data is necessary for establishing the dialogue structure for the TPR analysis. Furthermore, because the caretaker's linguistic knowledge is not in doubt, their data is also evaluated against the formal and functional benchmarks to provide a robustness test. For each D×N pair in a child utterance U, a language model undergoes CAC prompting with the context C, defined above, to elicit determiner preferences in identical context as the child.

This way, we obtain a model-generated corpus that parallels child-produced corpus: at each D×N site, the child has produced a particular determiner (*the* or *a*) while the model has produced a probabilistic distribution over them, which is then used to analytically derive its empirical predictions (Section 3.2). Both child and model data are evaluated with our formal and functional benchmarks.

Finally, we apply the benchmarks to the traditional forced-choice task by choosing determiners according to the model's maximum likelihood determiner (MLE) at each D×N site, in line with BLiMP and other such minimal pair tasks (Section 1). To that end, we also report the extent to which a model's majority determiner choices agree with children, whom previous work has shown to be adult-like. While we cannot exhaustively examine model output, strong divergence from child's usage may indict something amiss.

3.1 Model inference

We evaluate 45 language models spanning three architecture classes: masked language models (MLM), autoregressive (AR) language models, and encoder-decoder (seq2seq) models. These are the BabyLM baselines, winning submissions, and reference models publicly available as of March 2026. Full model metadata, including parameter counts, nominal training budgets, estimated total training exposure are reported in Appendix B, including whether the Manchester Corpus is known or likely to be included in training.

For each model, we apply CAC to obtain the logits of *the* and *a* at the site of each D×N combination in children's production, which are converted to a probabilistic distribution. Due to the differences in model architecture, there is some variation in model inference. For MLMs, we replace the target determiner with a single mask token and read off the logits at that position. Because all MLMs in our evaluation tokenize the determiners as single tokens, no multi-mask setup is required. Consistent with the child language studies, the phonological variant *an* is merged into *a* as a single choice.

For AR models, we explicitly construct and score three full candidate utterances: one with *the*, one with *a*, and one with *an* at the target position. Let $U^{(\mathrm{the})}$, $U^{(\mathrm{a})}$, and $U^{(\mathrm{an})}$ denote these variants. Each candidate is scored by its conditional probability under the shared discourse context:

$$p_\theta(U^{(k)} \mid C) = \prod_{t=1}^{T} p_\theta(w_t^{(k)} \mid C, w_{<t}^{(k)}),$$

where $k \in \{\mathrm{the}, \mathrm{a}, \mathrm{an}\}$. Again we merge $U^{(\mathrm{a})}$ and $U^{(\mathrm{an})}$ as a single choice.

For seq2seq models with span-corruption objectives, we use the standard sentinel-token interface and score three teacher-forced target sequences corresponding to the same determiner alternatives. The encoder input replaces the target determiner with a sentinel token, and the decoder assigns

$$p_\theta(y^{(k)} \mid X) = \prod_{t=1}^{T'} p_\theta(y_t^{(k)} \mid \mathrm{Enc}(X), y_{<t}^{(k)}),$$

where X is the encoder input with the determiner replaced by the sentinel and $k \in \{\mathrm{the}, \mathrm{a}, \mathrm{an}\}$. Finally, *a* and *an* are merged into a single choice.

3.2 Model evaluation

Because the CAC prompting produces a probabilistic distribution of *the* and *a* for each D×N site in the model-generated corpus, it is possible to analytically derive the expected values of D×N overlap and TPRs. We discuss these in turn.

Formal D×N productivity For each child corpus which has N unique noun types and S D×N combinations, CAC prompts a model to generate a distribution over (*the*, *a*) at each D×N site. Let k_n be the number of D×N sites for a noun type n, and $p_{n,i}$ be the probability of *the* at site i where $i = 1, 2, \ldots k_n$. The expected value of overlap $\widehat{O}$ for a model-generated corpus is the probability of a noun attested with both determiners averaged over all noun types:

$$\widehat{O} = \frac{1}{N} \sum_{n \in N} \hat{o}_n, \;\; \hat{o}_n = 1 - \prod_{i=1}^{k_n} p_{n,i} - \prod_{i=1}^{k_n} (1 - p_{n,i}) \tag{2}$$

The expected value of bias $\hat{b}$ for the corpus can also be derived analytically. It is computed at the noun level (cf. Eq 1): for each noun type n, we sum the expected counts of *the* and *a* across all k_n sites and take the larger sum as that noun's contribution to the preferred-determiner count:

$$\hat{b} = \frac{1}{S} \sum_{n \in N} \max \left(\sum_{i=1}^{k_n} p_{n,i}, \sum_{i=1}^{k_n} (1 - p_{n,i}) \right) \tag{3}$$

Together, N, S, and $\hat{b}$ determine the predicted overlap under a fully productive grammar (Eq. 8; Appendix A). These values will be compared against the expected empirical value $\widehat{O}$; the absence of significant difference supports productivity.

Determiner transitions. TPR calculation takes place over a pair of determiner transition $t = (C_t, M_t)$ sharing a topic noun, with C_t produced by the caretaker and M_t generated by the model. Here, C_t is either *the* or *a* but the model choice M_t is a probabilistic distribution (M_t^{the}, M_t^{a}). The expected value of TPR over all T transitions is:

$$\widehat{\mathrm{TPR}} = \frac{1}{T}\sum_{t\in T}\big(M_t^{the}P(C_t = \mathrm{a}) + M_t^{a}P(C_t = \mathrm{the})\big) \tag{4}$$

Model TPR is compared against children, caretakers, and a baseline from adult-to-adult conversations; see Section 4 for details. The comparison asks whether, in the statistical aggregate, a model uses determiners across conversational turns in a manner similar to human language users.

4 Results

We begin by testing human data against the benchmarks before turning to the models: both are summarized in Table 1.

Formal `D`×`N` productivity. In the Manchester corpus, children's mean empirical overlap was 0.251 (SD = 0.068) against the mean expected value of 0.242 (SD = 0.105), and caretakers' mean empirical overlap was 0.301 (SD = 0.045) against the mean expected value of 0.320 (SD = 0.070). A paired t-test found no significant difference between empirical and expected values for either group (children: $t(11) = 0.653, p = 0.527$; caretakers: $t(11) = -1.392, p = 0.191$), replicating the findings of Yang (2013) that determiners used by children (and adults) show productivity. Full dyad-level statistics are reported in Appendix C.

Note that the average bias value is remarkably similar between children (mean = 0.834, SD = 0.043) and caretakers (mean = 0.815, SD = 0.021): a paired t-test reveals no significant difference between children and caretakers ($t(11) = 1.798, p = 0.100$). One-sample t-tests also found that neither group's mean bias differed significantly from the universal baseline value of 0.82 established from COCA (Davies, 2008) (One sample t-test: children $t(11) = 1.168, p = 0.268$; caretakers $t(11) = -0.758, p = 0.464$). This confirms the similarity and stability of usage across English and in turn supports the robustness of the benchmark.

The caretakers' overlap values are considerably higher than the children's (paired t-test: t(11)=-3.616, p=0.004). Previous analyses (Pine and Martindale, 1996; Pine et al., 2013) regarded the difference as children's reduced productivity. However, the difference is due to adults talking more and having more opportunities for the nouns to be used with both determiners. This is reflected in the token/type ratio (S/N), which is higher for caretakers than for children (children: mean $S/N = 4.192$; caretakers: mean $S/N = 6.226$) and is strongly correlated with empirical overlap in both groups (children: $r(10) = .883, p < .001$; caretakers: $r(10) = .695, p = .012$).

Discourse-functional benchmark (TPR). We first established the TPR baseline of 0.215 by examining 1,615 possible determiner transitions from adult-to-adult dialogue in CHILDES. For the 12 children in the Manchester Corpus, the mean TPR when interacting with caretakers is 0.226 (SD = 0.050). For the 12 caretakers, the mean TPR when interacting with children is 0.200 (SD = 0.046), see Table 1. The children's TPR values do not significantly differ from their caretakers' on a paired t-test ($t = 2.061, p = 0.064$) even though caretakers sometimes repeat child speech verbatim. Neither children nor caretakers show difference from the adult-to-adult baseline (One sample t-test: children $t = 0.724$, $p = 0.484$; caretakers $t = 1.198$, $p = 0.256$). The statistical results suggest that the distribution of determiners in discourse is similar for children and adults alike, supporting earlier research that children use determiners appropriately.

LLM results. Table 1 summarizes representative model results compared to the human baselines across the two evaluation criteria, with the full results for all 45 models reported in Table D1. Figure 2 visualizes the representative models across all three diagnostic dimensions.

Only two (`ltg-bert-bnc` and `roberta-base`) models passed both the formal overlap test and the discourse-functional TPR test. Both models, however, are trained on much larger amount of the data (100M and 30B words) than available to 2-3 year olds. The `t5-base` model, trained on even more data (156B words), also passes the TPR test but fails the overlap test.

As shown in Table D1, 30 models satisfied the formal overlap test, including several models such

Table 1: Human and representative model results. Means and SDs are computed over the 12 Manchester child dyads. Checkmarks indicate no significant deviation from benchmark ($p > 0.05$). Full results are in Table D1.

Model	Arch.	Bias	Empirical overlap (SD)	Predicted overlap (SD)	TPR (SD)	DxN test	TPR test
Children	Human	0.834	0.251 (0.068)	0.242 (0.105)	0.226 (0.050)	✓ **(0.527)**	✓ **(0.484)**
Caretakers	Human	0.815	0.301 (0.045)	0.320 (0.070)	0.200 (0.046)	✓ **(0.191)**	✓ **(0.256)**
ltg-bert-bnc	MLM	0.783	0.314 (0.068)	0.287 (0.107)	0.225 (0.048)	✓ **(0.138)**	✓ **(0.505)**
roberta-base	MLM	0.776	0.323 (0.071)	0.292 (0.112)	0.239 (0.051)	✓ **(0.096)**	✓ **(0.142)**
t5-base	S2S	0.753	0.347 (0.062)	0.304 (0.102)	0.239 (0.042)	× (0.030)	✓ **(0.085)**
babylm-baseline-100m-gpt-bert-causal-focus	AR	0.782	0.314 (0.068)	0.288 (0.106)	0.247 (0.047)	✓ **(0.122)**	× (0.041)
gpt-bert-babylm-small	MLM	0.902	0.185 (0.054)	0.185 (0.091)	0.431 (0.087)	✓ **(0.997)**	× (<0.001)
opt-125m	AR	0.756	0.346 (0.067)	0.303 (0.104)	0.251 (0.044)	× (0.027)	× (0.018)
gpt2	AR	0.752	0.349 (0.069)	0.305 (0.104)	0.252 (0.042)	× (0.031)	× (0.012)
roberta-base-strict-2023	MLM	0.834	0.301 (0.067)	0.249 (0.086)	0.475 (0.062)	× (0.001)	× (<0.001)

as `ltgbert-10m-2024` and `gpt-bert-small` that were trained on an amount of data (10M words) comparable to that for 2-3 year olds. Setting aside four models whose outputs were degenerate due to always predicting the same determiner, this is an impressive result: Language models can learn the formal productivity of determiners and apply them interchangeably to nouns. However, none of these models pass the discourse-functional test, with their TPRs typically much higher: 0.433 for `ltgbert-10m-2024` and 0.438 for `gpt-bert-small`, compared to human speakers at 0.215. It is as if the models choose determiners stochastically with no regard for referential restrictions, as shown in the snippets of caretaker-child conversation.

Finally, we also evaluated the maximum likelihood determiner choices at each `D`×`N` site by the models. Doing so brings evaluation in line with classification tasks, and also allows us to measure model accuracy in comparison to children. Several models achieved high accuracy: for example, `babylm-baseline-100m-gpt-bert-causal-focus` and `babylm-baseline-100m-gpt-bert-mixed` match child choices at 0.898 and 0.889, respectively, yet neither passed the TPR test, and only narrowly satisfied the overlap test. This supports our earlier remark (Section 1) that classification is not a rigorous demonstration of adult-like knowledge.

5 Related Work

Productivity test The overlap measure has generated several quantitative assessments of child language. As noted earlier, they all require input data, which is available in very limited ways. For example, Alhama et al. (2024) mark determiner productivity as the recording session where children use both determiners with at least two unique nouns. That, however, does not rule out the possibility of them having memorized them from prior but unrecorded input data. Similarly, Meylan et al. (2017) estimate determiner probabilities for specific nouns, including the idiosyncratic *bathroom* and *bath*, to infer the degree of productivity. Due to the sparsity of the input data, the inference model fails to converge for the majority of cases under study. Valian et al. (2009) and Pine et al. (2013) compare children and caretakers directly and interpret a higher overlap value as greater productivity. But the overlap value itself can be misleading. As Yang (2013) observes, the overlap value of the Brown Corpus (Kučera and Francis, 1967), a collection of print materials, is lower than some children (Pine and Martindale, 1996) but professional writers clearly do not have a worse grammar than two-year-olds.

CAC and LLM inference CAC prompting adapts the cloze task (Taylor, 1953), which has previously served as a behavioral probe of language models across domains, from morphosyntactic analysis in context (Cotterell et al., 2018; McCarthy et al., 2019) to cross-modal brain decoding (Zou et al., 2022). CAC departs from prior cloze-style probes in two respects. First, it restricts the choice set to a closed grammatical alternation rather than an open vocabulary, motivated by recent large-scale comparisons showing that LM next-token distributions and human cloze completions are not interchangeable (Jacobs et al., 2024). Second, it is inspired by discourse cloze tasks in psycholinguistics (Chihara et al., 1977) and NLP (Paperno et al., 2016), rather than in the local sentence alone as in previous LLM work on determiners (Alhama et al., 2023; Fiandra et al., 2025). It accentuates the recognition that language must accurately fulfill communicative functions beyond the well-formedness of sentences.

The probability distributions elicited by CAC are then treated as direct behavioral evidence, similar to a framework for LM surprisal analysis. Surprisal theory (Levy, 2008) establishes that the negative

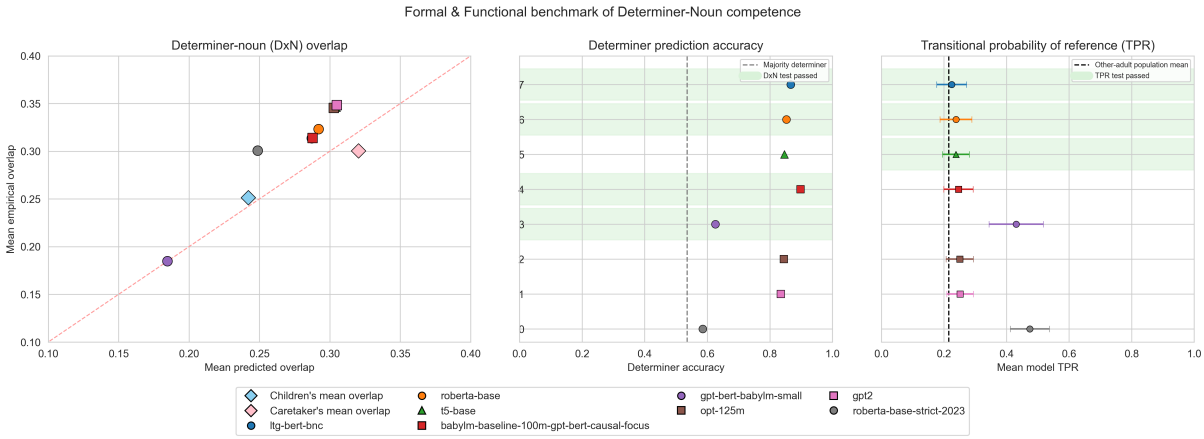

Figure 2: Formal and functional benchmarks for the representative models in Table 1. Left: mean empirical vs. predicted D×N overlap, with the dashed diagonal indicating identity and children/caretakers shown for reference. Middle: determiner prediction accuracy, with models passing the formal overlap test highlighted. Right: mean TPR across dyads, with the dashed line marking the adult population mean and models passing the TPR test highlighted.

log-probability of a word in context is an interpretable signal of the model's own linguistic expectations, a link that appears to hold across languages (Wilcox et al., 2023). Model probabilities may thus provide a more reliable measure of linguistic knowledge than verbalized or prompted outputs even in instruction-tuned models (Hu and Levy, 2023; Kauf et al., 2024).

6 Discussion and Conclusion

We introduce two quantitative benchmarks for language beyond classification that draw attention to the functional aspects of language that cannot be accurately assessed by well-formeness measures alone. Indeed, in the TPR analysis, both *the* and *a* are fully grammatical but the accurate choice must take discourse function and structure into account. We have shown that current language models trained on developmentally plausible data do not yet exhibit the knowledge of determiners that English-learning children demonstrate by age 2-3. Some, but but not all, models trained on much more data do meet the benchmarks but it is not clear what causes success or failure. The three-way dissociation between determiner prediction accuracy, formal D×N productivity, and discourse-functional determiner use shows that these are different facets of competence, and that optimizing for one provides no guarantee of acquiring the others.

Beyond the empirical findings, we also make a methodological contribution. The statistical benchmarks require no access to training data and are applicable to any language sample produced by children or models alike. The overlap test in particular has been adapted to study other linguistic phenomena in humans (Goldin-Meadow and Yang, 2017a) and can be extended to LLMs as well. CAC, the prompting method introduced to elicit model output suitable for these benchmarks, is not specific to determiners: it can be applied to any grammatical choice at an attested site in naturalistic dialogue.

More generally, our work calls for closer attention to child language research (Vázquez Martínez et al., 2023). Early acquisition of determiners serves as a checkpoint for LLMs. Other aspects of morphology and grammar are learned later with frequent errors (Ervin, 1964; Chomsky, 1969; Yang, 2016) and can also serve as a constraining factor: LLMs may also learn slowly with comparable errors. Finally, it is well known that the same element of grammar may be learned along different schedules across languages. For example, Italian- and Mandarin-learning children are adult-like in their usage of grammatical subjects around 2 (Valian, 1991; Wang et al., 1992) but English-learning children only manage to reach adult-level by 4 (Bloom, 1970). Similarly, children acquiring English go through years where a finite verb is substituted with an infinitive (e.g., *He eat cookies*) but children acquiring Spanish rarely do so (Legate and Yang, 2007). And so on. Such patterns of development, including variation within and across languages, have been amply documented and should shed light on the status of LLMs as cognitive models, especially with the BabyLM project expanding the range of languages under investigation.

Limitations

Scope of the productivity test. The formal productivity test has additional constraints that limit its generality. First, the closed-class category must have exactly two members: the mathematical derivation of expected overlap is tractable for a binary category such as {*the*, *a*} but becomes intractable with additional members such as *this* and *that*. Second, the test assumes that the two categories combine in a fully interchangeable, statistically independent way. This characterizes determiners and nouns well, but not all grammatical combinations have this property. Not all transitive verbs passivize (*John resembles Bill* cannot become **Bill was resembled by John*), not all dative verbs participate in both the double-object and *to*-dative constructions, and so on. CAC can be deployed for such phenomena, but the formal productivity test cannot straightforwardly follow.

Corpus contamination. Many of the evaluated models were likely trained on corpora that include the Manchester data. Despite our efforts to control for this using raw probability weights, contamination cannot be eliminated entirely, and results for models with known or likely Manchester exposure should be interpreted with appropriate caution.

Compute constraints. A single inference run through all 45 models required approximately 14 days of wall-clock time, which prevented evaluation of models at the 1B parameter scale and above. Larger models, and in particular frontier instruction-tuned and RLHF-trained systems, present an additional methodological consideration beyond scale: instruction tuning and reinforcement learning from human feedback optimize for conversational behavior that may interact with the CAC prompting interface in ways that differ from models trained on language modeling objectives alone. Whether such models meet the formal and functional criteria presented here is the topic of future work.

References

Raquel G. Alhama, Ruthe Foushee, Dan Byrne, Allyson Ettinger, Afra Alishahi, and Susan Goldin-Meadow. 2024. Using computational modeling to validate the onset of productive determiner–noun combinations in English-learning children. *Proceedings of the National Academy of Sciences*, 121(50):e2316527121.

Raquel G. Alhama, Ruthe Foushee, Daniel Byrne, Allyson Ettinger, Susan Goldin-Meadow, and Afra Alishahi. 2023. Linguistic Productivity: the Case of Determiners in English. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 330–343, Nusa Dua, Bali. Association for Computational Linguistics.

Jean Berko. 1958. The child's learning of English morphology. *Word*, 14(2–3):150–177.

Lois Bloom. 1970. *Language development: Form and function in emerging grammar*. MIT Press, Cambridge, MA.

Melissa Bowerman. 1982. Reorganizational process in lexical and syntactic development. In Eric Wanner and Lila R. Gleitman, editors, *Language acquisition: The state of the art*, pages 319–346. Cambridge University Press, New York.

Roger Brown. 1973. *First Language: The Early Stages*. Harvard University Press, Cambridge.

Tetsuro Chihara, John Oller, Kelley Weaver, and Mary Anne Chavez-Oller. 1977. Are cloze items sensitive to constraints across sentences? 1. *Language learning*, 27(1):63–70.

Carol Chomsky. 1969. *The acquisition of syntax in children from 5 to 10*. MIT Press, Cambridge, MA.

Noam Chomsky. 1955. The logical structure of linguistic theory. Manuscript, Harvard University.

Noam Chomsky. 1957. *Syntactic structures*. Mouton, The Hague.

Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Jaap Jumelet, Tal Linzen, Aaron Mueller, Suchir Salhan, Raj Sanjay Shah, Alex Warstadt, and Ethan Gotlieb Wilcox. 2026. BabyLM Turns 4 and Goes Multilingual: Call for Papers for the 2026 BabyLM Workshop. *arXiv preprint*. ArXiv:2602.20092 [cs].

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.

Mark Davies. 2008. Corpus of Contemporary American English (COCA): 410+ million words, 1990-present. Available at http://www.americancorpus.org/.

Jill G De Villiers and Peter A De Villiers. 1974. Competence and performance in child language: Are children really competent to judge? *Journal of Child Language*, 1(1):11–22.

Susan M. Ervin. 1964. Imitation and structural change in children's language. In Eric H. Lenneberg, editor, *New directions in the study of child language*, pages 163–189. MIT Press, Cambridge.

Larry Fenson, Philip S Dale, J Steven Reznick, Elizabeth Bates, Donna J Thal, Stephen J Pethick, Michael Tomasello, Carolyn B Mervis, and Joan Stiles. 1994. Variability in early communicative development. *Monographs of the Society for Research in Child Development*, pages i–185.

Olivia La Fiandra, Nathalie Fernandez Echeverri, Patrick Shafto, and Naomi H. Feldman. 2025. Large Language Models and Children Have Different Learning Trajectories in Determiner Acquisition. In *Proceedings of the First BabyLM Workshop*, pages 100–108, Suzhou, China. Association for Computational Linguistics.

Jill Gilkerson, Jeffrey A Richards, Steven F Warren, Judith K Montgomery, Charles R Greenwood, D Kimbrough Oller, John HL Hansen, and Terrance D Paul. 2017. Mapping the early language environment using all-day recordings and automated analysis. *American journal of speech-language pathology*, 26(2):248–265.

Talmy Givón. 1983. *Topic continuity in discourse*. John Benjamins Publishing Company.

Lila Gleitman and Charles Yang. 2022. Two-year-olds' referential determiners in discourse. In *Boston Unversity Conference on Language Development*.

Susan Goldin-Meadow and Charles Yang. 2017a. Statistical evidence that a child can create a combinatorial linguistic system without external linguistic input: Implications for language evolution. *Neuroscience and Biobehavioral Reviews*, 81(Part B):150 – 157.

Susan Goldin-Meadow and Charles Yang. 2017b. Statistical evidence that a child can create a combinatorial linguistic system without external linguistic input: Implications for language evolution. *Neuroscience & Biobehavioral Reviews*, 81:150–157.

Zellig S. Harris. 1951. *Methods in structural linguistics.* University of Chicago Press, Chicago.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060.

Cassandra L. Jacobs, Loïc Grobol, and Alvin Tsang. 2024. Large-scale cloze evaluation reveals that token prediction tasks are neither lexically nor semantically aligned. *arXiv preprint*. Version Number: 2.

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A Ivanova. 2024. Log probabilities are a reliable estimate of semantic plausibility in base and instruction-tuned language models. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 263–277, Miami, Florida, US. Association for Computational Linguistics.

Erwin Komen, Rosanne Hebing, Ans Van Kemenade, and Bettelou Los. 2014. Quantifying information structure change in english. In *Information structure and syntactic change in Germanic and Romance languages*, pages 81–110. John Benjamins Publishing Company, Amsterdam.

Henry Kučera and W. Nelson Francis. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence.

Julie A. Legate and Charles Yang. 2007. Morphosyntactic learning and the development of tense. *Language Acquisition*, 14(3):315–344.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Christopher Lyons. 1999. *Definiteness*. Cambridge University Press, Cambridge.

Brian MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk*, 3rd edition. Lawrence Erlbaum, Mahwah, NJ.

Michael Maratsos. 1976. *The use of definite and indefinite reference in young children*. Cambridge University Press, London.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2023. How Much Do Language Models Copy From Their Training Data? Evaluating Linguistic Novelty in Text Generation Using RAVEN. *Transactions of the Association for Computational Linguistics*, 11:652–670. Place: Cambridge, MA.

William Merrill, Noah A. Smith, and Yanai Elazar. 2024. Evaluating n-Gram Novelty of Language Models Using Rusty-DAWG. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language*

Processing, pages 14459–14473, Miami, Florida, USA. Association for Computational Linguistics.

Stephan C. Meylan, Michael C. Frank, Brandon C. Roy, and Roger Levy. 2017. The Emergence of an Abstract Grammatical Category in Children's Early Speech. *Psychological Science*, 28(2):181–192.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc-Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1525–1534.

Julian M Pine, Daniel Freudenthal, Grzegorz Krajewski, and Fernand Gobet. 2013. Do young children have adult-like syntactic categories? Zipf's law and the case of the determiner. *Cognition*, 127(3):345–360.

Julian M Pine and Helen Martindale. 1996. Syntactic categories in the speech of young children: The case of the determiner. *Journal of child language*, 23(2):369–395.

Margot Isabella Rozendaal and Anne Edith Baker. 2008. A cross-linguistic investigation of the acquisition of the pragmatics of indefinite and definite reference in two-year-olds. *Journal of Child Language*, 35(4):773–807.

Rushen Shi. 2023. Functional categories, determiners, prosody and early child grammar. In D. Armoskaite, S. Armoskaite, and M Wiltschko, editors, *Oxford handbook of determiners*. Oxford University Press.

Dan Slobin. 1985. *The crosslinguistic study of language acquisition*. Lawrence Erlbaum, Mahwah, NJ.

Jon Sprouse, Beracah Yankama, Sagar Indurkhya, Sandiway Fong, and Robert C Berwick. 2018. Colorless green ideas do sleep furiously: gradient acceptability and the nature of the grammar. *The Linguistic Review*, 35(3):575–599.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Herbert S Terrace, Laura-Ann Petitto, Richard J Sanders, and Thomas G Bever. 1979. Can an ape create a sentence? *Science*, 206(4421):891–902.

Anna Theakston, Elena Lieven, Julian Pine, and Caroline Rowland. 2001. The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of child language*, 28:127–52.

Rosalind Thornton. 2021. Judgments of acceptability, truth, and felicity in child language. *The Cambridge handbook of experimental syntax (Cambridge handbooks in language and linguistics)*, pages 394–420.

Michael Tomasello. 2000. Do young children have adult syntactic competence? *Cognition*, 74(3):209–253.

Virginia Valian. 1986. Syntactic categories in the speech of young children. *Developmental Psychology*, 22(4):562.

Virginia Valian. 1991. Syntactic subjects in the early speech of American and Italian children. *Cognition*, 40(1):21–81.

Virginia Valian, Stephanie Solt, and John Stewart. 2009. Abstract categories or limited-scope formulae? The case of children's determiners. *Journal of Child Language*, 36(4):743–778.

Héctor Vázquez Martínez, Annika Lea Heuser, Charles Yang, and Jordan Kodner. 2023. Evaluating Neural Language Models as Cognitive Models of Language Acquisition. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 48–64, Singapore. Association for Computational Linguistics.

Qi Wang, Diane Lillo-Martin, Catherine T Best, and Andrea Levitt. 1992. Null subject versus null object: Some evidence from the acquisition of Chinese and English. *Language Acquisition*, 2(3):221–254.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Charles Yang. 2013. Ontogeny and phylogeny of language. *Proceedings of the National Academy of Sciences*, 110(16):6324–6327.

Charles Yang. 2016. *The Price of Linguistic Productivity: How Children Learn to Break the Rules of Language*. The MIT Press.

George Kingsley Zipf. 1949. *Human Behavior And The Principle Of Least Effort*. Addison-Wesley Press.

Shuxian Zou, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2022. Cross-modal cloze task: A new task to brain-to-word decoding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 648–657, Dublin, Ireland. Association for Computational Linguistics.

A Derivation of the formal D×N productivity test

This appendix gives the derivation of the expected-overlap formula used in the main text for the formal D×N productivity test. Following Yang (2013), the

derivation assumes two statistical properties of the corpus: (i) singular noun frequencies follow an Zipf's Law, and (ii) nouns may exhibit an aggregate bias toward one determiner over the other.

Noun rank probabilities. Let S be the number of D×N combinations involving N unique noun types. Assuming nouns follow Zipf's Law, the noun of rank r has probability

$$p_r = \frac{1}{r^a H_N} \tag{5}$$

$$\text{where} \quad H_N = \sum_{i=1}^{N} \frac{1}{i}. \tag{6}$$

Determiner bias. See Eq. 1 in the main text.

Expected overlap for a noun of rank r. For a noun of rank r, the probability that it is never observed at all in S draws is

$$(1 - p_r)^S.$$

Now consider the event that the noun is observed, but only with one determiner. Under the aggregate bias b, the probability that a token is either not this noun or is this noun with its favored determiner is

$$1 - (1 - b)p_r = bp_r + (1 - p_r),$$

so the probability that the noun never appears with its disfavored determiner is

$$(bp_r + 1 - p_r)^S.$$

Subtracting the probability that the noun is never observed at all gives the probability that it appears only with its favored determiner:

$$(bp_r + 1 - p_r)^S - (1 - p_r)^S.$$

By the same reasoning, the probability that the noun appears only with its disfavored determiner is

$$((1 - b)p_r + 1 - p_r)^S - (1 - p_r)^S.$$

A noun contributes to overlap if it appears with both determiners at least once. Therefore, the expected overlap contribution of the noun of rank r is one minus the probabilities of the three complementary events: never observed, observed only with the favored determiner, and observed only with the disfavored determiner. This yields

$$\begin{aligned} E_r = 1 &- (1 - p_r)^S \\ &- \left[(bp_r + 1 - p_r)^S - (1 - p_r)^S\right] \\ &- \left[((1 - b)p_r + 1 - p_r)^S - (1 - p_r)^S\right]. \end{aligned} \tag{7}$$

Expected overlap for the sample. The expected overlap for the full sample is the mean of these noun-level expectations over all noun ranks:

$$E[S] = \frac{1}{N} \sum_{r=1}^{N} E_r. \tag{8}$$

Empirical comparison. The empirical overlap statistic used in the main text is the proportion of noun types that occur with both determiners among all noun types that occur with either:

$$\text{empirical} = \frac{1}{|N|} \sum_{n \in N} \mathbb{1}\left[\forall d \in \{\textit{the}, a\}, C_{d \times n} > 0\right] \tag{9}$$

The formal productivity test compares this empirical value with the expected value in Equation 8. If the two do not differ significantly, the sample is treated as consistent with a fully productive D×N system.

Scope of the test. Although we apply the test here to English determiner–noun combinations, the logic is more general. As discussed by Yang (2013) and Goldin-Meadow and Yang (2017b), the same framework can be extended to other two-way combinatorial systems in which one category is small and closed, the other is larger and approximately Zipfian, and the question is whether the observed overlap is compatible with productive combination. Yang (2013) also used it to reject productivty in Nim Chimpsky's ASL signs, providing independent support for distributional analysis of videos by his trainers (Terrace et al., 1979).

B Evaluated Models

Tables B1–B3 list the language models evaluated in this study together with metadata relevant to the interpretation of our determiner experiments. Model names are given as `HuggingFace` repository paths where applicable. For each model, we report its parameter count, nominal training-data budget, the estimated total number of tokens seen during training, and whether the Manchester corpus is known or likely to be included in pretraining. We distinguish between nominal dataset size and total training exposure because several models were trained for multiple passes over relatively small corpora, so the total number of tokens seen during training exceeds the training budget allowance.

Column descriptions. **Dataset size** denotes the nominal size of the training corpus or benchmark

track. **Total tokens seen** reports our compiled estimate of total training exposure. **Manchester in pretraining?** indicates whether the evaluation corpus is documented as included in pretraining, likely included, excluded, or unknown. We observed Manchester material directly in the BabyLM Strict and Strict-Small training corpora; accordingly, models trained on these datasets are marked as having included the Manchester corpus during pretraining.

Column descriptions. **Dataset size** denotes the nominal size of the training corpus or benchmark track. **Total tokens seen** reports our compiled estimate of total training exposure. **Manchester in pretraining?** indicates whether the Manchester corpus is known to be included in pretraining (*yes*), confirmed absent (*no*), likely but not independently confirmed (*possibly*), or not determinable from available documentation (*unknown*). Models trained on the BabyLM Strict (100M) or Strict-Small (10M) datasets are marked *yes*: we confirmed that both tracks contain Manchester corpus material. Models trained on other corpora are marked *yes* where independent confirmation was possible, *possibly* where the training data composition makes inclusion likely but could not be verified, and *unknown* where no information about corpus composition was available.

C Human validation on Manchester

As a validation of our preprocessing pipeline, we first applied the same extraction and overlap-computation procedure used throughout the paper to the human Manchester data. Singular noun phrases were identified automatically using the spaCy's `en-core-web-sm` noun chunk extractor (Honnibal and Johnson, 2015), after which we computed determiner–noun counts, empirical overlap, predicted overlap, and TPR for each child and caretaker sample. (Our results and conclusion hold when using the annotation provided by CHILDES.) We apply this same preprocessing pipeline to the language models evaluated in the main experiments.

The resulting raw overlap and TPR statistics are shown in Table C1. These values closely reproduce the formal productivity pattern reported by Yang (2013): in the Manchester corpus, children's empirical `D×N` overlap is statistically indistinguishable from their predicted overlap values ($p = 0.527$). This result is consistent with the productivity expected under a fully productive system, and caretaker samples show the same general profile ($p = 0.191$) despite differences in sample size and vocabulary. The same table also shows that both children and caretakers are statistically indistinguishable from the adult population TPR average (children: $p = 0.484$; caretakers: $p = 0.256$). In this sense, our results align with the broader use of the overlap framework in Goldin-Meadow and Yang (2017b) and validate the TPR benchmark as well.

We include these statistics here for two reasons. First, they provide a sanity check on our implementation of both overlap analyses. Second, they corroborate our Manchester preprocessing pipeline as the basis for model-side evaluation, since the pipeline recovers the established human baseline on both the formal and functional benchmarks before being applied to BabyLM systems.

D Full model results

Table D1 reports the full model-by-model results for all evaluated systems. It combines the formal `D×N` productivity results and the functional discourse Transitional Probability of Reference (TPR) results into a single table.

For each model, we report mean empirical overlap and mean predicted overlap across the 12 Manchester child dyads, with standard deviations in parentheses, together with determiner bias, determiner accuracy, the outcome of the formal `D×N` test, mean TPR, and the outcome of the TPR test. In the two test columns, the symbol indicates whether the model passed the corresponding criterion, and the exact p-value is given in parentheses in the same cell.

Models with non-significant differences between empirical and predicted overlap are treated as consistent with formal `D×N` productivity. Models with non-significant differences under the TPR comparison are considered to have achieved discourse-appropriate functional usage of determiners. The main-text summary table is a compact subset of this full table.

Although four models with `bias`≥ 0.98 appear to pass the formal `D×N` test, this is an artifact of degenerate determiner predictions. Concretely, all three `BabyBERTa` variants and `babylm-10m-gpt-bert-mixed (MLM)` overwhelmingly predict the same determiner ('a(n)'), as evidenced by the bias value being close to 1,

Table B1: Masked language models evaluated in this study. Full Hugging Face repository paths are provided here; later tables use shortened model names for readability.

Repository / family	Variant	Params.	Dataset size	Total tokens seen	Manchester?
phueb/BabyBERTa	-1	5M	5M words	50M words	No
	-2	5M	5M words	50M words	No
	-3	5M	5M words	50M words	No
lgcharpe/ELC_BERT	_small_baby_10M	24M	10M words	~127B	Yes
	_baby_100M	98M	100M words	~270B	Yes
ltg/ltg-bert	-bnc	98M	100M words	124B	No
	-babylm	98M	100M words	124B	Yes
	-10m-2024	30M	10M words	Unknown	Possibly
	-100m-2024	119M	100M words	Unknown	Yes
BabyLM-community/ babylm-baseline-10m-gpt-bert	-masked-focus	31M	10M words	100M words	Yes
	-mixed	31M	10M words	100M words	Yes
BabyLM-community/ babylm-baseline-100m-gpt-bert	-masked-focus	120M	100M words	1B words	Yes
	-mixed	120M	100M words	1B words	Yes
roberta	-base	125M	~30B	~2T	No
	-large	355M	~30–50B tokens	~2T	No
nyu-mll/roberta-med-small	-1M-1	45M	1M words	~26B tokens	No
	-1M-2	45M	1M words	~2.6B tokens	No
	-1M-3	45M	1M words	~8.1B tokens	No
nyu-mll/roberta-base	-10M-1	125M	10M words	~10B tokens	No
	-10M-2	125M	10M words	~2.6B tokens	No
	-10M-3	125M	10M words	~8.1B tokens	No
	-100M-1	125M	100M words	~26B tokens	No
	-100M-2	125M	100M words	~33B tokens	No
	-100M-3	125M	100M words	~8.1B tokens	No
	-1B-1	125M	1B words	~26B tokens	No
	-1B-2	125M	1B words	~33B tokens	No
	-1B-3	125M	1B words	~520B tokens	No
babylm/roberta-base-strict	-2023	125M	100M words	2B words	Yes
	-small-2023	125M	10M words	200M words	Yes

Table B2: Autoregressive models evaluated in this study. Full Hugging Face repository paths are provided here; later tables use shortened model names for readability.

Repository / family	Variant	Params.	Dataset size	Total tokens seen	Manchester?
ltg/gpt-bert-babylm	-small	30M	10M words	15.6B	Possibly
	-base	119M	100M words	31.25B	Possibly
BabyLM-community/ babylm-baseline-10m-gpt-bert	-causal-focus	31M	10M words	100M words	Yes
	-mixed	31M	10M words	100M words	Yes
BabyLM-community/ babylm-baseline-100m-gpt-bert	-causal-focus	120M	100M words	1B words	Yes
	-mixed	120M	100M words	1B words	Yes
timinar/baby-llama	-58m	58M	10M words	Unknown	Yes
babylm/babyllama	-10m-2024	58M	10M words	~100M words	Yes
	-100m-2024	58M	100M words	~1B words	Yes
facebook/opt	-125m	125M	180B	~180B tokens	Unknown
babylm/opt-125m-strict	-2023	125M	100M words	2B words	Yes
	-small-2023	125M	10M words	200M words	Yes
openai-community/gpt	2	124M	~8B tokens	~10B tokens	Unknown
BabyLM-community/ babylm-baseline	-10m-gpt2	124M	10M words	100M words	Yes
	-100m-gpt2	124M	100M words	1B words	Yes

Table B3: Encoder–decoder models evaluated in this study. Full Hugging Face repository paths are provided here; later tables use shortened model names for readability.

Repository / family	Variant	Params.	Dataset size	Total tokens seen	Manchester?
`google-t5/t5`	`-base`	220M	156B words	34B tokens	Possibly
`babylm/t5-base-strict`	`-2023`	220M	100M words	2B words	Yes
	`-small-2023`	220M	10M words	200M words	Possibly

Table C1: Types (N), tokens (S), determiner bias score, token/type ratio (r), and observed (empirical) versus predicted overlap values for 12 children and their corresponding caretakers in the Manchester corpus. Singular noun phrases were extracted automatically using spaCy's noun chunker. At the group level, children's empirical overlap does not differ significantly from predicted overlap ($p = 0.527$), and the same is true for caretakers ($p = 0.191$), reproducing the formal productivity pattern reported in prior work and validating the preprocessing pipeline used throughout the paper.

Dyad	Speaker	$\|N\|$	$\|S\|$	Bias	Empirical	Predicted	n_{TPR}	TPR
Gail	Child	316	863	0.868	0.193	0.148	224	0.241
	Caretaker	838	3578	0.839	0.251	0.217	286	0.192
Dominic	Child	123	323	0.904	0.130	0.132	172	0.221
	Caretaker	539	4205	0.791	0.282	0.417	188	0.261
Becky	Child	364	1385	0.846	0.255	0.212	438	0.194
	Caretaker	592	3519	0.831	0.326	0.304	548	0.197
Liz	Child	312	1291	0.862	0.269	0.217	373	0.166
	Caretaker	619	3022	0.838	0.265	0.252	410	0.193
Carl	Child	407	3684	0.770	0.386	0.494	910	0.260
	Caretaker	516	3669	0.794	0.355	0.389	1077	0.208
Joel	Child	336	1020	0.889	0.179	0.144	331	0.145
	Caretaker	819	3550	0.836	0.220	0.223	445	0.133
Ruth	Child	203	747	0.798	0.246	0.258	279	0.272
	Caretaker	707	4668	0.796	0.310	0.355	372	0.183
Aran	Child	376	1635	0.813	0.258	0.263	802	0.261
	Caretaker	1072	8272	0.807	0.312	0.372	1079	0.152
Anne	Child	317	1170	0.815	0.290	0.233	500	0.168
	Caretaker	720	6083	0.831	0.368	0.386	657	0.158
John	Child	333	1615	0.807	0.279	0.296	473	0.307
	Caretaker	740	3876	0.789	0.341	0.301	526	0.276
Nicole	Child	195	492	0.860	0.210	0.152	234	0.214
	Caretaker	833	4372	0.836	0.262	0.260	280	0.179
Warren	Child	397	2314	0.782	0.320	0.355	826	0.262
	Caretaker	854	6080	0.797	0.316	0.367	944	0.265

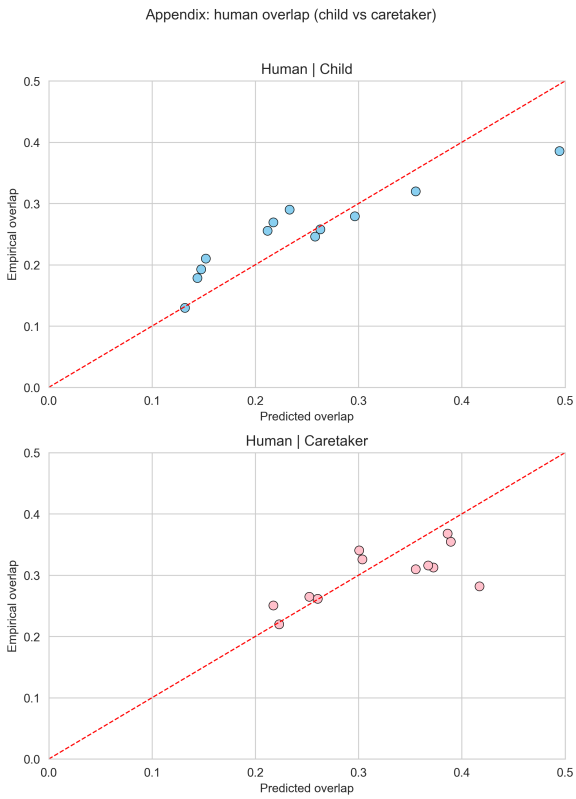

Figure C1: Full formal D×N overlap results for 12 children (top) and their corresponding caretakers (bottom) in the Manchester corpus. No statistically significant difference is found between the empirical and predicted overlap scores (paired t-test $p = 0.527$ children and $p = 0.191$ for caretakers), meeting the formal D×N overlap criterion.

yielding trivially small overlap values rather than substantive evidence of productivity.

E Full results figures

Figure E1 visualizes the formal D×N overlap results for the full set of evaluated models, with children and caretakers shown as human reference points. Figure E2 shows determiner prediction accuracy, with models that pass the formal overlap test highlighted. Figure E3 shows mean TPR by model with standard-deviation whiskers, alongside the human child benchmark and models that pass the TPR criterion.

Table D1: Full model-by-model results for all evaluated systems. The table reports determiner bias, mean empirical and predicted D×N overlap across the 12 Manchester child dyads, mean TPR, the outcomes of the formal D×N and TPR tests, and determiner accuracy is the agreement with children when the MLE determiner is chosen by the model. Checkmarks indicate passing ($p > 0.05$); exact p-values are shown in parentheses.

Model	Arch.	bias	empirical (SD)	predicted (SD)	TPR (SD)	D×N test	TPR test	Accuracy
Children	Human	0.834	0.251 (0.068)	0.242 (0.105)	0.226 (0.050)	✓ **(0.527)**	✓ **(0.484)**	–
Caretakers	Human	0.815	0.301 (0.045)	0.320 (0.070)	0.200 (0.046)	✓ **(0.191)**	✓ **(0.256)**	–
BabyBERTa-1	MLM	0.996	0.014 (0.008)	0.014 (0.007)	0.497 (0.100)	✓ **(0.057)**	× (<0.001)	0.580
BabyBERTa-2	MLM	0.999	0.005 (0.003)	0.005 (0.003)	0.497 (0.100)	✓ **(0.445)**	× (<0.001)	0.580
BabyBERTa-3	MLM	0.999	0.006 (0.004)	0.006 (0.004)	0.497 (0.100)	✓ **(0.863)**	× (<0.001)	0.580
BabyLM-2026-Baseline-GPT2-Strict	AR	0.701	0.381 (0.062)	0.327 (0.096)	0.393 (0.024)	× (0.007)	× (<0.001)	0.653
BabyLM-2026-Baseline-GPT2-Strict-Small	AR	0.718	0.375 (0.064)	0.321 (0.102)	0.379 (0.036)	× (0.010)	× (<0.001)	0.700
ELC$_BERT_baby_100M$	MLM	0.799	0.296 (0.081)	0.277 (0.118)	0.299 (0.058)	✓ **(0.308)**	× (<0.001)	0.781
ELC$_BERT_small_baby_10M$	MLM	0.949	0.143 (0.042)	0.116 (0.040)	0.498 (0.090)	× (<0.001)	× (<0.001)	0.580
baby-llama-58m	AR	0.741	0.352 (0.061)	0.309 (0.100)	0.271 (0.046)	× (0.025)	× (0.002)	0.817
babyllama-100m-2024	AR	0.775	0.319 (0.070)	0.292 (0.108)	0.256 (0.050)	✓ **(0.126)**	× (0.017)	0.885
babyllama-10m-2024	AR	0.768	0.327 (0.067)	0.296 (0.107)	0.257 (0.048)	✓ **(0.094)**	× (0.012)	0.870
babylm-baseline-100m-gpt-bert-causal-focus	AR	0.782	0.314 (0.068)	0.288 (0.106)	0.247 (0.047)	✓ **(0.122)**	× (0.041)	0.898
babylm-baseline-100m-gpt-bert-masked-focus	MLM	0.980	0.062 (0.052)	0.054 (0.041)	0.489 (0.094)	× (0.043)	× (<0.001)	0.582
babylm-baseline-100m-gpt-bert-mixed	AR	0.777	0.319 (0.069)	0.291 (0.107)	0.251 (0.046)	✓ **(0.113)**	× (0.023)	0.889
babylm-baseline-100m-gpt-bert-mixed	MLM	0.976	0.074 (0.038)	0.067 (0.035)	0.485 (0.095)	× (0.013)	× (<0.001)	0.581
babylm-baseline-100m-gpt2	AR	0.729	0.366 (0.075)	0.317 (0.107)	0.405 (0.041)	× (0.013)	× (<0.001)	0.726
babylm-baseline-10m-gpt-bert-causal-focus	AR	0.762	0.334 (0.067)	0.300 (0.109)	0.267 (0.048)	✓ **(0.072)**	× (0.003)	0.861
babylm-baseline-10m-gpt-bert-masked-focus	MLM	0.957	0.118 (0.054)	0.101 (0.046)	0.488 (0.092)	× (0.003)	× (<0.001)	0.581
babylm-baseline-10m-gpt-bert-mixed	AR	0.756	0.340 (0.069)	0.303 (0.107)	0.266 (0.043)	✓ **(0.060)**	× (0.002)	0.856
babylm-baseline-10m-gpt-bert-mixed	MLM	0.984	0.052 (0.035)	0.049 (0.031)	0.487 (0.096)	✓ **(0.128)**	× (<0.001)	0.582
babylm-baseline-10m-gpt2	AR	0.740	0.348 (0.065)	0.311 (0.108)	0.389 (0.047)	✓ **(0.066)**	× (<0.001)	0.725
gpt-bert-babylm-base	MLM	0.916	0.175 (0.059)	0.168 (0.081)	0.441 (0.081)	✓ **(0.560)**	× (<0.001)	0.606
gpt-bert-babylm-small	MLM	0.902	0.185 (0.054)	0.185 (0.091)	0.431 (0.087)	✓ **(0.997)**	× (<0.001)	0.626
gpt2	AR	0.752	0.349 (0.069)	0.305 (0.104)	0.252 (0.042)	× (0.031)	× (0.012)	0.835
ltg-bert-babylm	MLM	0.797	0.297 (0.082)	0.278 (0.119)	0.314 (0.059)	✓ **(0.289)**	× (<0.001)	0.774
ltg-bert-bnc	MLM	0.783	0.314 (0.068)	0.287 (0.107)	0.225 (0.048)	✓ **(0.138)**	✓ **(0.505)**	0.866
ltgbert-100m-2024	MLM	0.739	0.352 (0.071)	0.311 (0.110)	0.308 (0.057)	✓ **(0.055)**	× (<0.001)	0.776
ltgbert-10m-2024	MLM	0.792	0.287 (0.055)	0.280 (0.106)	0.433 (0.059)	✓ **(0.714)**	× (<0.001)	0.640
opt-125m	AR	0.756	0.346 (0.067)	0.303 (0.104)	0.251 (0.044)	× (0.027)	× (0.018)	0.845
opt-125m-strict-2023	AR	0.812	0.299 (0.060)	0.263 (0.071)	0.430 (0.068)	× (0.001)	× (<0.001)	0.500
opt-125m-strict-small-2023	AR	0.745	0.354 (0.067)	0.308 (0.098)	0.368 (0.030)	× (0.014)	× (<0.001)	0.697
roberta-base	MLM	0.776	0.323 (0.071)	0.292 (0.112)	0.239 (0.051)	✓ **(0.096)**	✓ **(0.142)**	0.853
roberta-base-100M-1	MLM	0.756	0.343 (0.072)	0.303 (0.115)	0.267 (0.050)	✓ **(0.052)**	× (0.004)	0.787
roberta-base-100M-2	MLM	0.752	0.343 (0.076)	0.305 (0.117)	0.285 (0.053)	✓ **(0.070)**	× (<0.001)	0.796
roberta-base-100M-3	MLM	0.748	0.353 (0.076)	0.308 (0.115)	0.282 (0.051)	× (0.030)	× (<0.001)	0.780
roberta-base-10M-1	MLM	0.862	0.225 (0.076)	0.227 (0.113)	0.367 (0.068)	✓ **(0.907)**	× (<0.001)	0.702
roberta-base-10M-2	MLM	0.823	0.278 (0.082)	0.260 (0.120)	0.338 (0.055)	✓ **(0.328)**	× (<0.001)	0.746
roberta-base-10M-3	MLM	0.836	0.267 (0.081)	0.250 (0.111)	0.347 (0.065)	✓ **(0.296)**	× (<0.001)	0.729
roberta-base-1B-1	MLM	0.770	0.331 (0.077)	0.295 (0.119)	0.278 (0.055)	✓ **(0.072)**	× (0.002)	0.786
roberta-base-1B-2	MLM	0.747	0.351 (0.076)	0.308 (0.115)	0.283 (0.048)	× (0.031)	× (<0.001)	0.797
roberta-base-1B-3	MLM	0.775	0.324 (0.076)	0.292 (0.117)	0.285 (0.059)	✓ **(0.101)**	× (0.002)	0.805
roberta-base-strict-2023	MLM	0.834	0.301 (0.067)	0.249 (0.086)	0.475 (0.062)	× (0.001)	× (<0.001)	0.585
roberta-base-strict-small-2023	MLM	0.777	0.355 (0.074)	0.290 (0.102)	0.487 (0.056)	× (0.002)	× (<0.001)	0.584
roberta-large	MLM	0.766	0.331 (0.071)	0.297 (0.116)	0.270 (0.054)	✓ **(0.111)**	× (0.005)	0.832
roberta-med-small-1M-1	MLM	0.754	0.331 (0.059)	0.303 (0.099)	0.475 (0.042)	✓ **(0.149)**	× (<0.001)	0.601
roberta-med-small-1M-2	MLM	0.693	0.390 (0.070)	0.332 (0.107)	0.459 (0.019)	× (0.010)	× (<0.001)	0.587
roberta-med-small-1M-3	MLM	0.787	0.300 (0.061)	0.284 (0.100)	0.476 (0.051)	✓ **(0.363)**	× (<0.001)	0.594
t5-base	S2S	0.753	0.347 (0.062)	0.304 (0.102)	0.239 (0.042)	× (0.030)	✓ **(0.085)**	0.846
t5-base-strict-2023	S2S	0.786	0.315 (0.083)	0.286 (0.114)	0.360 (0.067)	✓ **(0.106)**	× (<0.001)	0.748
t5-base-strict-small-2023	S2S	0.657	0.420 (0.067)	0.341 (0.090)	0.483 (0.037)	× (<0.001)	× (<0.001)	0.436

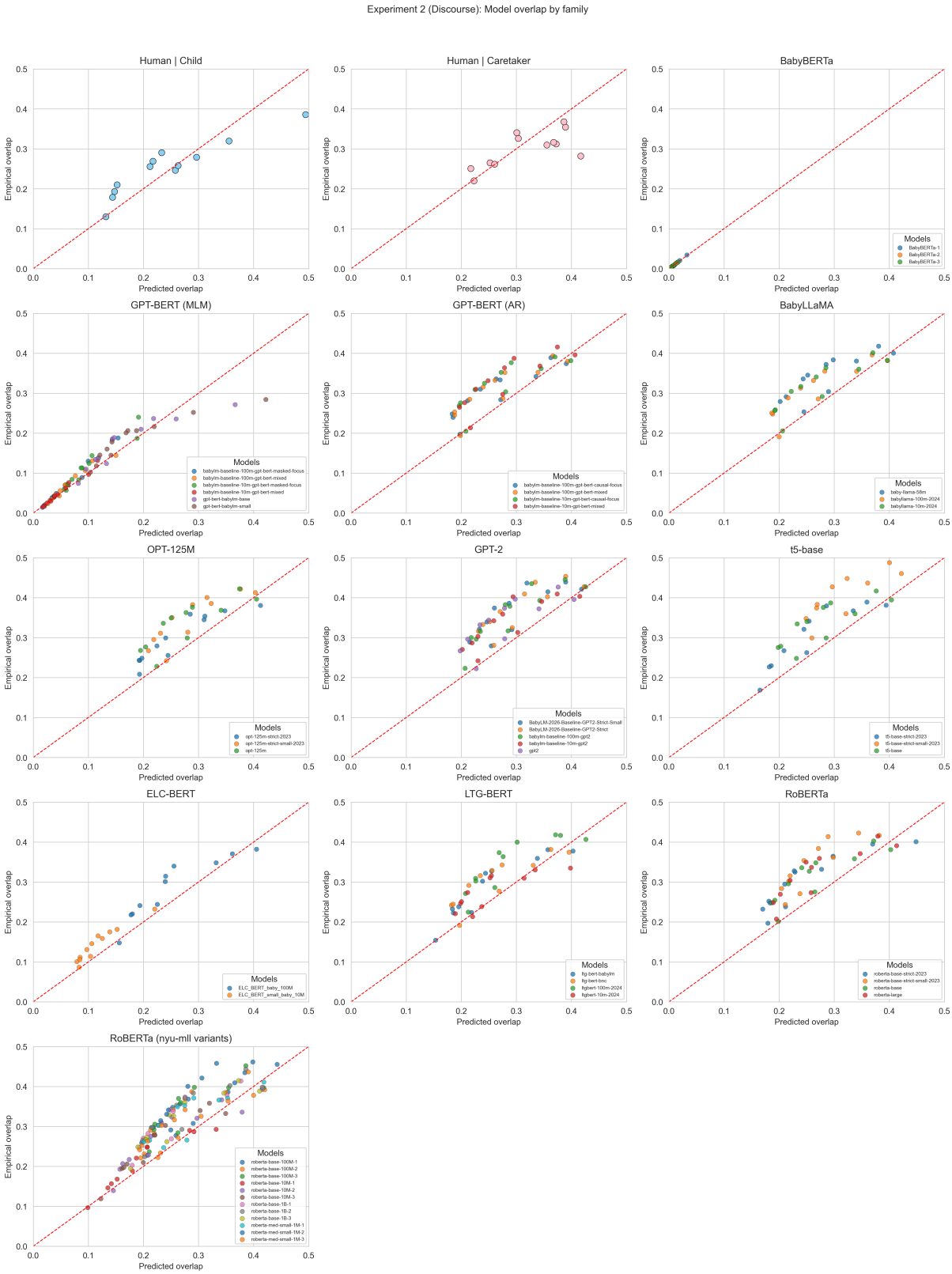

Figure E1: Full formal D×N overlap results for all evaluated models. Each panel shows empirical overlap against predicted overlap for a model family, with the diagonal line indicating equality between observed and predicted overlap. The child and caretaker Manchester dyads are included as human reference panels.

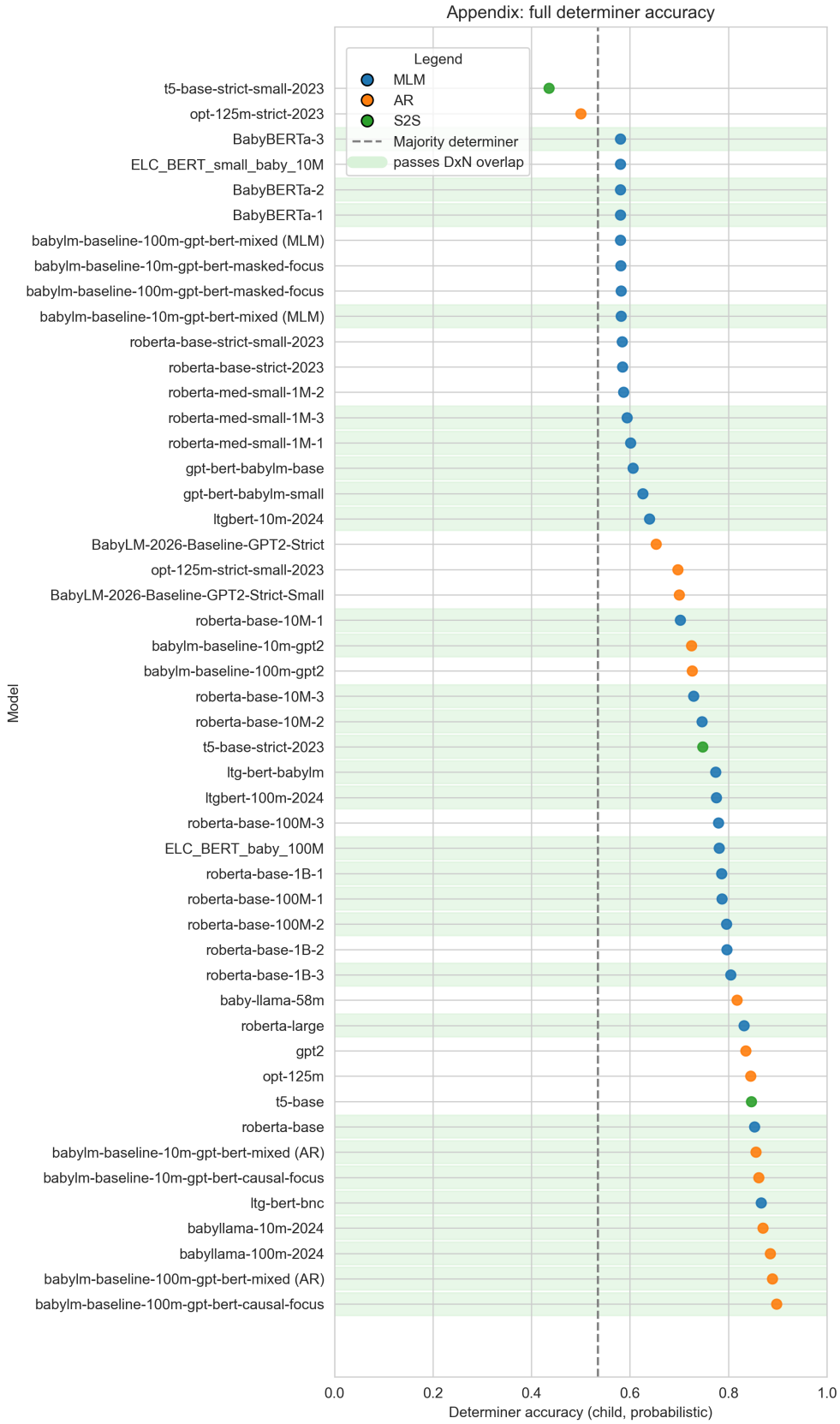

Figure E2: Full determiner prediction accuracy results for all evaluated models. The dashed vertical line indicates the accuracy of a majority-determiner baseline. Models that pass the formal D×N overlap test are highlighted in green.

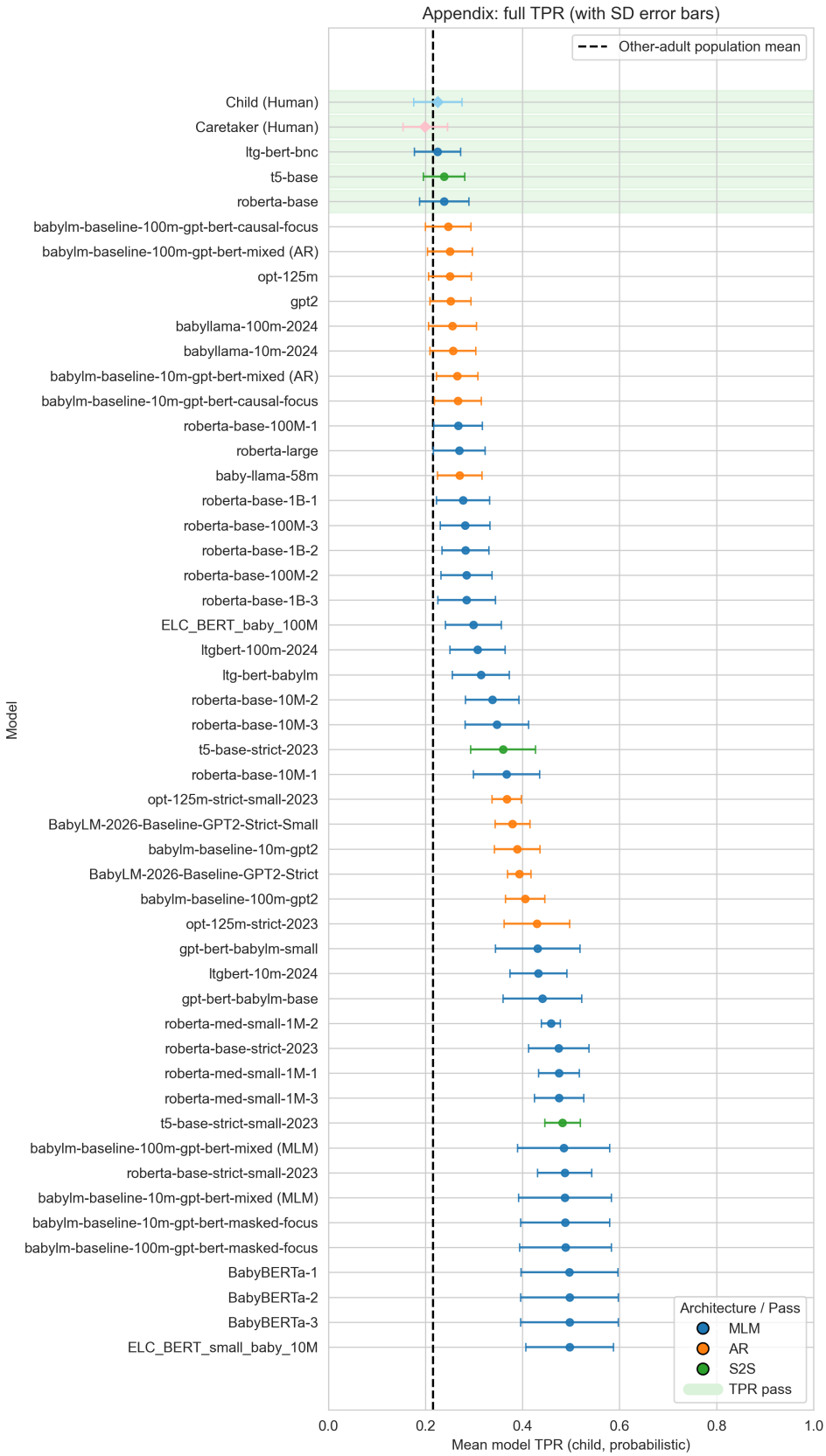

Figure E3: Full Transitional Probability of Reference (TPR) results for all evaluated models, with the 12 Manchester corpus children and corresponding caretakers at the top for reference. Points show mean TPR and whiskers show standard deviations across 12 dyads. The dashed vertical line indicates the adult population mean TPR, and models that pass the TPR test are highlighted in green.