Do BabyLMs *Wanna* Learn *Wanna* Contraction? On the Learnability without Language-Specific Bias

Kangsan Noh
Korea University
kasan1998@korea.ac.kr

Sanghoun Song
Korea University
sanghoun@korea.ac.kr

Abstract

This study investigates whether the grammatical constraints on *wanna* contraction—a phenomenon traditionally cited as evidence for innate linguistic knowledge—can be learned via BabyLMs, which are designed to reflect cognitively plausible learning conditions. Two datasets were constructed from the CHILDES corpus, varying in embedded verb frequency (high vs. low) and grammaticality, and contrasting grammatical instances (object extraction contexts) with ungrammatical ones (subject extraction contexts) of *wanna* contractions. Using surprisal as a metric, we evaluated 24 BabyLMs from the 2024 BabyLM Challenge alongside four standard models, including BERT and GPT-2. While the standard models performed with near-perfect consistency, the BabyLMs showed modest but meaningful sensitivity, particularly those trained on larger datasets and tested on high-frequency *wanna* instances. In particular, only encoder-based BabyLMs captured the grammatical constraint, with `babylm24_MLSM` exhibiting consistent performance. Nonetheless, our findings provide evidence for limited and conditional learnability of *wanna* contraction by artificial learners under cognitively realistic input conditions.

1 Introduction

This study examines whether language models designed to reflect cognitively plausible learning conditions can recognize grammatical constraints. Specifically, we focus on the learnability of the constraint on *wanna* contraction, a phenomenon traditionally cited as evidence for innate linguistic knowledge in language acquisition (Chomsky, 1977; Crain and Thornton, 1998). However, more recent experimental research has challenged this traditional view, suggesting that learning mechanisms, particularly those sensitive to distributional information, may contribute to explaining the constraint on *wanna* contraction (Zukowski and Larsen, 2011; Getz, 2019). Extending this line of research, Noh et al. (2024) tested standard language models for their sensitivity to the constraint and found that the models generally responded appropriately. Building on this body of prior work, we employ BabyLMs to investigate the extent to which the constraint on *wanna* contraction can be learned by artificial learners that lack significant advantages over human learners in the amount of available data.

Since the release of the Transformer architecture (Vaswani et al., 2017), language models have played a transformative role in Natural Language Processing and have increasingly influenced research in linguistics and cognitive science (Ambridge and Blything, 2024; Niu et al., 2024). The extent to which these artificial learners are comparable to humans has been the subject of ongoing debate, eliciting both optimistic and skeptical perspectives of their contributions (Blank, 2023). Skeptics argue that although language models may be revolutionary and impressive from an engineering standpoint, they offer, if any, limited insights for linguistic theory (Chomsky et al., 2023; Katzir, 2023; Fox and Katzir, 2024). Specifically, it is often pointed out that their enormous parameter size and reliance on vast amounts of training data make them fundamentally different from how humans acquire language. In contrast, proponents argue that, despite these limitations, language models offer valuable insights into long-standing debates on language acquisition, the poverty of the stimulus, and the learnability of grammatical constraints (Warstadt and Bowman, 2022; Kallens et al., 2023; Piantadosi, 2024).

In light of the debate over the relevance of language models to research on language acquisition, this study addresses two central research questions.

Q1: Is the grammatical constraint on *wanna* contraction learnable not only by standard language models but also by BabyLMs, which are designed

Findings of the Association for Computational Linguistics: ACL 2026, pages 11361–11372

to avoid significant advantages over human learners in terms of input size? To address the first question, we compared the performance of BabyLMs and standard models on both grammatical and ungrammatical instances of *wanna* contraction.

Q2: To what extent does the frequency of remnant input patterns influence BabyLMs' processing of *wanna* contractions? To address the second, we evaluated the model performance using two datasets that varied in the frequency of embedded verbs following *wanna*.

2 Background

2.1 BERTology

The advent of Transformer-based language models has given rise to a research program informally known as *BERTology* — the systematic investigation of what linguistic knowledge these models encode and how they do so (Rogers et al., 2020). A central methodological insight driving this line of work is to treat language models as psycholinguistic subjects (Linzen et al., 2016; Futrell et al., 2019). Researchers design controlled test sets modeled after experimental paradigms in psycholinguistics, evaluating models by examining whether they assign higher probabilities to grammatical sentences than to minimally differing ungrammatical counterparts. Benchmarks such as BLiMP (Warstadt et al., 2020) and SyntaxGym (Gauthier et al., 2020) operationalize grammaticality contrasts as minimal pairs, enabling systematic comparison of models across a wide range of syntactic phenomena.

A recurring question in this body of work is whether grammatical generalizations that have traditionally been attributed to innate linguistic knowledge are in fact learnable from distributional input alone. For instance, Wilcox et al. (2022) investigated the learnability of filler-gap dependencies and the island constraints on them, finding that language models acquire not only the basic contingency between fillers and gaps, but also its hierarchical and unbounded properties, suggesting that at least some grammatical constraints may be learnable from distributional input without necessarily invoking innate linguistic biases.

2.2 *Wanna* Contraction

We selected *wanna* contraction as the target phenomenon to evaluate BabyLMs' ability to process grammatical constraints. Specifically, *wanna* contraction involves the reduction of the verb *want* and the infinitival marker *to* into the contracted form *wanna*, as illustrated in the pair in (1).

(1) a. Who do you want to kiss?
    b. Who do you wanna kiss?

Notably, *wanna* contraction is not uniformly permissible. There are contexts in which it is blocked (Lakoff, 1970), as illustrated in (2) and (3).

(2) a. Who do you want t_i to come tomorrow?
    b. *Who do you wanna come tomorrow?

(3) a. Who do you want to see t_i tomorrow?
    b. Who do you wanna see tomorrow?

Traditionally, the *wh*-trace account has been the leading explanation for the ungrammaticality of certain instances of *wanna* contraction, rooted in the Universal Grammar framework and based on the assumption of an invisible *wh*-trace (Lightfoot, 1976; Chomsky, 1977; Chomsky and Lasnik, 1977; Rotenberg, 1978). Under the *wh*-trace account, contraction is not permitted in subject-extraction contexts like (2), where a *wh*-trace intervenes between *want* and *to*. By contrast, in object-extraction contexts like (3), the *wh*-trace appears after the embedded verb and does not prevent contraction.

This constraint on *wanna* contraction has been a central topic of debate among theoretical and experimental linguists (Lakoff, 1970; Lightfoot, 1976; Chomsky, 1977; Postal and Pullum, 1978, 1982; Crain and Thornton, 1998; Boas, 2004; Zukowski and Larsen, 2011; Ito, 2018; Goodall, 2021; Hwang, 2023). A traditional analysis argues that this constraint cannot be learned from the input alone, as children are rarely exposed to sufficient evidence indicating when the contraction is ungrammatical. Thus, this analysis appeals to Universal Grammar to account for how learners acquire knowledge of when the contraction is permitted and when it is not.

Note that the goal of this study is not to support or refute traditional syntactic analyses of *wanna* contraction. Rather, this study focuses on the distinction between grammatical (object extraction) and ungrammatical (subject extraction) instances of *wanna* contraction, investigating whether BabyLMs are sensitive to the syntactic environments in which contraction is disallowed. The constraint on *wanna* contraction thus serves as a probe for syntactic knowledge acquired from linguistic input alone, without language-specific predispositions.

2.3 BabyLM Challenge

Most language models are typically trained using massive amounts of data. According to the BabyLM Challenge website, BERT (Devlin et al., 2019) was trained on approximately 3 billion words, RoBERTa (Liu et al., 2019) on 30 billion words, and GPT-3 (Brown et al., 2020) on as many as 200 billion words. Launched in 2023, the BabyLM Challenge promoted the development of cognitively plausible models of language acquisition. These models aim to be efficient in terms of parameter size and training data, while also being relevant to research on human language learning (Warstadt et al., 2023a). To achieve this goal, the challenge imposes developmentally realistic training conditions by limiting models to corpora of 10 million or 100 million words, a scale roughly comparable to the linguistic input a child receives by the age of 13 years. Therefore, scaling down the training data size is essential for advancing cognitively plausible language modeling.

The BabyLM Challenge has undergone two iterations. The first challenge, held in 2023, attracted more than 30 submissions that introduced novel training strategies and model architectures (Warstadt et al., 2023b). Building on this momentum, the second challenge in 2024 (Hu et al., 2024) expanded its scope by introducing new evaluation tracks incorporating multimodal and multilingual inputs, as well as more challenging benchmarks. The BabyLM Challenge focused on improving data efficiency and promoting cognitive plausibility to better reflect human language acquisition. Building on these prior studies, we used BabyLMs as our primary subjects.

3 Methods

3.1 Test Material

The test sentences used in this study were based on Getz (2019), who considered the frequency of embedded verbs in *wanna*. The frequency classification of embedded verbs in Getz (2019) was based on their occurrences in the CHILDES Parental Corpus. According to this criterion, the embedded verbs were categorized into three frequency groups: ‘low’, ‘medium’, and ‘high.’ By controlling for frequency, Getz (2019) aimed to determine whether sensitivity to ungrammatical cases of *wanna* contraction was influenced by the frequency of the embedded verbs. The results indicate that there is indeed a frequency effect. When induced to produce sentences containing *wanna*, children were less likely to generate ungrammatical instances when the embedded verbs were of higher frequency.

Consequently, we controlled the frequency as a factor in constructing our test sentences. We first downloaded and analyzed the `train_100M.zip` and `train_10M.zip` files from the BabyLM Challenge OSF repository by extracting the instances of *want to* and *wanna* usage, as presented in Table 1.[1]

Table 1: Occurrences in `train_100.zip` and `train_10.zip` from the BabyLM Challenge.

Corpus	Size	WANT TO	WANNA
BNC Spoken	100M	5,831	1,901
	10M	681	206
CHILDES	100M	9,057	59,085
	10M	845	6,256
Gutenberg	100M	7,727	0
	10M	857	0
Open Subtitles	100M	25,815	6,708
	10M	2,646	835
Simple Wiki	100M	971	50
	10M	88	5
Switchboard	100M	894	0
	10M	88	0

Note that *wanna* occurred more frequently than *want to* only in the CHILDES data. Considering that *wanna* appeared most frequently in this subset, we employed a regular expression to extract all words immediately following *wanna* from the CHILDES portion of the 100M dataset.[2] Then we manually identified the embedded verbs, excluding erroneous tokens and non-verbal expressions such as nouns or adjectives. Consequently, we obtained embedded verbs following *wanna* and their occurrences. From this list, we selected two high-frequency and two low-frequency verb pairs. Although selecting a larger set of verb pairs would have been ideal, we selected four in total, as these were the ones that clearly instantiated the contrast between object extraction and subject extraction. To ensure a clear contrast, we adopted a binary classification of frequency (‘high’ vs. ‘low’) rather than a tripartite classification (‘high’, ‘medium’, ‘low’). The selected pairs of embedded verbs are presented in Table 2.

[1]The BabyLM Challenge OSF repository is available at: https://osf.io/ryjfm/

[2]Specifically, we used the following regular expression: `\bwanna\s+(\w+)`

Table 2: Selected pairs of embedded verbs from `childes_100M.train`.

Pair	Embedded Verb	Sentence Type	Raw Frequency
High (A)	*go*	Subject-Extraction	5,727
	take	Object-Extraction	1,358
High (B)	*come*	Subject-Extraction	1,040
	see	Object-Extraction	3,284
Low (A)	*return*	Subject-Extraction	1
	lead	Object-Extraction	1
Low (B)	*apologize*	Subject-Extraction	1
	congratulate	Object-Extraction	1

In the list, *go* was the most frequent embedded verb, appearing 5,727 times. Along with *go*, we selected *take*, *come*, and *see*, each of which appeared over 1,000 times as an embedded verb. For a clear contrast, we selected *return*, *lead*, *apologize*, and *congratulate*, each of which occurred only once in the embedded position. Although it would have been preferable to include a larger set of embedded verbs, this was constrained by the highly skewed frequency distribution of verbs that occur with *wanna* in the CHILDES corpus, consistent with Zipf's law. To ensure statistical reliability, we restricted our analysis to manually selected verb pairs.

Using these pairs, we constructed high-frequency and low-frequency datasets, each consisting of 100 sentence pairs. In the high-frequency dataset, 50 pairs were constructed using *go/take* and another 50 using *come/see*. Similarly, in the low-frequency dataset, 50 pairs were constructed using *return/lead* and another 50 using *apologize/congratulate*. Example pairs are listed in Table 3. As in previous studies, our test sentences also contrast grammatical instances (object-extraction contexts) with ungrammatical instances (subject-extraction contexts). In each pair, we used different adverbial phrases to create distinct situational contexts (e.g., *to the station*, *airport*, *concert*; *after the meeting*, *debate*, *contest*, etc.). All test sentences shared a common temporal condition: the word *tomorrow* appeared at the end. The rationale for using *tomorrow* is explained in Section 3.2.

3.2 Surprisal

Surprisal serves as a metric for assessing sentence processing difficulty (Hale, 2001, 2016). For example, Futrell et al. (2019) define the surprisal of a target word x_i as shown in (4).

(4) $S(x_i) = -\log_2(p(x_i \mid h_{i-1}))$

As illustrated in (4), the surprisal of a word x_i, denoted as $S(x_i)$, is calculated by taking the logarithm of the reciprocal of its predicted probability, conditioned on the model's previous hidden state h_{i-1}. When a word is assigned a lower probability, its surprisal value becomes higher, indicating that the model finds it less predictable. Specifically, sentences containing unexpected or less frequent structures tend to yield higher surprisal scores.

In the BabyLM Challenge, the model performance was assessed using standardized evaluation suites, such as BLiMP (Warstadt et al., 2020), EWoK (Ivanova et al., 2024), and GLUE/SuperGLUE (Wang et al., 2018, 2019). However, considering the goal of the present study, which is to investigate whether BabyLMs can capture the grammatical constraints on *wanna* contraction, we adopt surprisal as our primary evaluation metric, following established psycholinguistic methods for assessing grammatical knowledge. Surprisal provides a gradient measure of processing difficulty, making it particularly well-suited for testing fine-grained grammatical sensitivity.

Based on this property of surprisal, we predict that the surprisal value for an ungrammatical instance of *wanna* contraction, such as (5a), is higher than that for a grammatical instance, such as (5b).

(5) a. *Who do you wanna go to the station tomorrow?
 b. Who do you wanna take to the station tomorrow?

Based on this assumption, we computed surprisal values in two ways: using encoder-based and decoder-based models. For the encoder-based models, we masked the final word *tomorrow* in each sentence, as shown in (6).

(6) a. Who do you wanna go to the station [MASK]?
 b. Who do you wanna take to the station [MASK]?

The logic behind this design is as follows: When processing sentence (6a), the models should recognize it as ungrammatical upon reaching the final

Table 3: Exemplary pairs from the high-frequency and low-frequency datasets.

Pair	Embedded Verb	Sentence Type	Sentence
High (A)	*go* *take*	Subject-Extraction Object-Extraction	*Who do you wanna go to the station tomorrow? Who do you wanna take to the station tomorrow?
High (B)	*come* *see*	Subject-Extraction Object-Extraction	*Who do you wanna come after the meeting tomorrow? Who do you wanna see after the meeting tomorrow?
Low (A)	*return* *lead*	Subject-Extraction Object-Extraction	*Who do you wanna return to the station tomorrow? Who do you wanna lead to the station tomorrow?
Low (B)	*apologize* *congratulate*	Subject-Extraction Object-Extraction	*Who do you wanna apologize after the meeting tomorrow? Who do you wanna congratulate after the meeting tomorrow?

word *tomorrow*, as there is no available position for object extraction—an essential condition for *wanna* contraction to be grammatical. In contrast, in sentence (6b), *tomorrow* should not be surprising, as the sentence already includes a valid position for object extraction following the embedded verb *take*. This reasoning applies to both encoder- and decoder-based language models. In encoder-based models, *tomorrow* serves as the target for the [MASK] tokens, whereas in decoder-based models, it is the final token to be predicted. [3]

Despite architectural differences, the core contrast remains the same: a model that successfully captures the grammatical constraint on *wanna* contraction should assign higher surprisal to (6a) than to (6b), since only the latter provides a grammatically licit environment for contraction. In (6a), the subject-extraction context blocks contraction, whereas in (6b), the presence of an object-extraction site licenses it. In short, a sensitive model should show increased surprisal for (6a) but not for (6b).

3.3 Baseline Experiment

We compared the performance of two groups of language models — standard language models and BabyLMs — across two experiments. In the baseline experiment, we used standard language models trained without restrictions on training data. In the main experiment, we used BabyLMs, which are trained with cognitive plausibility in mind and exposed to a more moderate amount of data than their large-scale counterparts.

The purpose of the baseline experiment is twofold. First, it provides a benchmark against which the performance of BabyLMs in the main experiment can be evaluated. Second, it serves to validate the test sentences: if standard language models perform well on these materials, any limitations in BabyLMs performance reflect genuine learning constraints rather than flaws in the test set itself.

In the baseline experiment, we used standard BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019) models to establish a point of comparison. Specifically, we deployed four pretrained models: BERT-base-uncased, BERT-large-uncased, GPT-2, and GPT-2-medium. BERT-base-uncased and BERT-large-uncased are trained using a masked language modeling objective and are based on the encoder block of the Transformer architecture. In contrast, GPT-2 and GPT-2-medium are autoregressive models that rely on the decoder block of the Transformer architecture. As the BabyLM Challenge includes both encoder- and decoder-based models, we deployed both BERT and GPT-2 variants to establish a baseline for comparison with the performance of BabyLMs.

3.4 Main Experiment

In the main experiment, we employed 20 BabyLMs listed on the 2024 BabyLM Leaderboard , including 10 models from the Strict Track (trained on 100 million words) and 10 from the Strict-Small Track (trained on 10 million words). Table 4 and Table 5 present the submitted and baseline models from the Strict and Strict-Small Tracks, respectively. In the model selection process, we included publicly available models on Hugging Face, each accessible through the corresponding model card. Although, ideally, all models submitted to the BabyLM Challenge must be evaluated, we selected 10 models per track to ensure the scope of the study is manageable. In addition to the submitted models, we also evaluated the baseline models, including two from the Strict Track and two from the Strict-Small Track. Specifically, the encoder-only LTG-BERT (Samuel et al., 2023) and the decoder-only BabyLlama (Timiryasov and Tastet, 2023) were used.

[3] The code and data are publicly available at `https://github.com/San0727/wanna-babylm`.

Table 4: BabyLMs tested in the main experiment (Strict Track).

Type	Models	References
Submission	antlm-bert-ntp_mlm-100m	Yu et al. (2024)
	babble-strict-competition-entry	Goriely et al. (2024)
	babylm24_LSM_strict	Berend (2024)
	babylm24_LSM015_strict	Berend (2024)
	babylm24_MLSM_strict	Berend (2024)
	BERTtime-Stories-100m	Theodoropoulos et al. (2024)
	grapheme-llama	Bunzeck et al. (2025)
	phoneme-llama	Bunzeck et al. (2025)
	RoBERTa-strict-newELI5-baseline	Lucas et al. (2024)
	RoBERTa-strict-newELI5-curriculumMasking	Lucas et al. (2024)
Baseline	BabyLlama_Baseline	Timiryasov and Tastet (2023)
	LTG-BERT_Baseline	Samuel et al. (2023)

Table 5: BabyLMs tested in the main experiment (Strict-Small Track).

Type	Models	References
Submission	antlm-bert-ntp_mlm-10m	Yu et al. (2024)
	BabyLlama-2-run1	Tastet and Timiryasov (2024)
	BabyLlama-2-run2	Tastet and Timiryasov (2024)
	babylm24_LSM_strict-small	Berend (2024)
	babylm24_LSM015_strict-small	Berend (2024)
	babylm24_MLSM_strict-small	Berend (2024)
	DeBaby-fullcontr	Edman (2024)
	DeBaby-halfcontr	Edman (2024)
	ELC-ParserBERT	Behr (2024)
	RoBERTa-strict-small_newELI5_baseline	Lucas et al. (2024)
Baseline	BabyLlama_Baseline	Timiryasov and Tastet (2023)
	LTG-BERT_Baseline	Samuel et al. (2023)

These two models were selected as baselines for the 2024 BabyLM Challenge, because they were the winning submissions from the previous year's challenge (Hu et al., 2024).

4 Results

4.1 Baseline Experiment

In the baseline experiment, we tested four standard language models that were not restricted in terms of training data. We made two predictions: First, the performance of the standard models would be near perfect, considering their extensive training data. Second, we expected the models to perform better on the high-frequency dataset than on the low-frequency dataset as the embedded verbs in the former occur more frequently with *wanna*.

The results of the paired t-tests indicate that our first prediction was largely correct, with only GPT-2 failing to show sensitivity to ungrammatical instances. The findings also showed that the standard models performed well on both the high-frequency and low-frequency datasets. Under both frequency conditions, the BERT models consistently exhibited significant differences. For the high-frequency dataset, both BERT-base-uncased (t = –12.233, p < .001) and BERT-large-uncased (t = –21.218, p < .001) exhibited strong effects, a pattern replicated in the low-frequency dataset (t = –6.640 and –15.680, respectively, with p < .001). Among the GPT-2 models, GPT-2-medium showed significant surprisal effects in both datasets (high-frequency: t = –4.981, p < .001; low-frequency: t = –9.123, p < .001), whereas GPT-2 failed to reach significance under either condition (high-frequency: t = –0.796, p = 0.428; low-frequency: t = –0.032, p = 0.975).

4.2 Main Experiment

In the main experiment, we tested 20 BabyLMs submitted to the 2024 BabyLM Challenge: 10 models trained on 100 million words (Strict Track) and 10 models trained on 10 million words (Strict-Small Track). Additionally, we tested four baseline models from the 2024 BabyLM Challenge: two from the Strict Track and two from the Strict-Small Track. Thus, a total of 24 models were tested.

As in the baseline experiment, we conducted paired t-tests to examine whether the BabyLMs submitted to the Strict and Strict-Small Track distinguished grammatical instances of *wanna* contraction from ungrammatical ones. We made two predictions. First, we expected that more models

Table 6: BabyLMs showing significant sensitivity to *wanna* contraction constraints. '—' indicates no significant sensitivity.

Track	Model	High-Freq	Low-Freq
Strict	babylm24_LSM_strict	-22.228 ($<.001$)	—
	babylm24_LSM015_strict	-20.530 ($<.001$)	—
	babylm24_MLSM_strict	-31.982 ($<.001$)	-20.164 ($<.001$)
	BERTtime-Stories-100m	-10.385 ($<.001$)	—
	LTG-BERT_Baseline	-18.973 ($<.001$)	-10.765 ($<.001$)
	RoBERTa-strict-newELI5-curriculumMasking	—	-9.419 ($<.001$)
Strict-Small	babylm24_LSM_strict-small	-13.031 ($<.001$)	—
	babylm24_LSM015_strict-small	-10.965 ($<.001$)	—
	babylm24_MLSM_strict-small	-17.229 ($<.001$)	—
	ELC-ParserBERT	—	-2.478 ($<.05$)

from the Strict Track would exhibit sensitivity to ungrammatical instances of *wanna* contraction than those from the Strict-Small Track, considering that the former were trained on ten times more data. Second, we anticipated more models would show sensitivity on the high-frequency dataset than on the low-frequency dataset, as the embedded verbs in the former occur more frequently with *wanna*.

The overall results are summarized in Table 6, providing empirical support for both of our predictions. More models from the Strict Track exhibited sensitivity than those from the Strict-Small Track, and sensitivity was more prevalent on the high-frequency dataset, highlighting the influence of the frequency effect. Notably, encoder-based models accounted for all cases of significant sensitivity, while none of the decoder-based models reached significance under any condition. This result was consistent across both the Strict and Strict-Small Tracks, suggesting that the pattern is robust regardless of training data size. This suggests that the training objective may be one of the factors influencing sensitivity to the constraint, a point we return to in the discussion.

5 Discussion

5.1 Performance Analysis of BabyLMs

The results showed that only the encoder-based BabyLMs exhibited sensitivity to ungrammatical instances of *wanna* contractions. Notably, babylm24_MLSM (Berend, 2024) consistently performed well, except for the Strict-Small Track with low-frequency data. In contrast to the standard Masked Language Modeling (MLM), which predicts missing words, the MLSM predicts abstract semantic concepts, which encourages deeper language understanding and stronger downstream performance (Berend, 2024). This distinction may be particularly relevant for the task at hand: determining whether *wanna* contraction is grammatical requires sensitivity to the semantic role of the embedded verb — specifically, whether it can take an object. In object-extraction contexts, the embedded verb takes an object that has been displaced, licensing the contraction, whereas in subject-extraction contexts, no such object position is available. By predicting abstract semantic concepts rather than surface word forms, MLSM may be better positioned to capture the argument structure of embedded verbs, which in turn facilitates the distinction between grammatical and ungrammatical instances of *wanna* contraction.

Although our experimental design was motivated by the characteristics of decoder-based models, which process inputs in a strictly unidirectional fashion, the prominent success of encoder-based models should not be interpreted as evidence that decoder-based models are inherently disadvantaged. In our setup, the masked token *tomorrow* appears at the final position of the sentence, which means that encoder-based models cannot attend to any right context beyond the mask, effectively neutralizing their bidirectional advantage. Nonetheless, the fact that encoder-based models consistently outperformed their decoder-based counterparts suggests that the masked language modeling objective itself may be particularly well-suited for tasks requiring sensitivity to syntactic constraints, as it encourages sensitivity to long-distance syntactic dependencies of the kind directly implicated in *wanna* contraction. The success of encoder-based models thus appears to reflect the inherent strength of the masked language modeling objective in capturing subtle grammatical constraints, rather than bidirectional processing per se — though we do not exclude the possibility that other architectural differences between encoder- and decoder-based models may also play a role.

5.2 Learnability of *Wanna* Contraction

At the heart of the ongoing debate over *wanna* contraction lies the key question: Are grammatical constraints inherent to an innate linguistic system? Proponents of linguistic nativism have traditionally attributed this constraint to Universal Grammar (Chomsky, 1977; Crain and Thornton, 1998), while others have questioned the role of such innate linguistic knowledge (Zukowski and Larsen, 2011; Getz, 2019). Therefore, *wanna* contraction serves as an ideal test case for examining the learnability of grammatical constraint by artificial learners that seemingly lack innate linguistic knowledge. First, the contracted form *wanna* is not frequently attested in corpus data, as demonstrated in this study. Second, ungrammatical (subject extraction) contexts, where the contraction is disallowed, are also relatively rare compared to grammatical (object extraction) contexts (Zukowski and Larsen, 2011). These two features of *wanna* contraction provide a compelling ground on which to test the validity of the Poverty of the Stimulus argument (Chomsky, 1980). As skeptics of language models argue, if what language models do is simply to learn heuristics from statistical patterns, it should be difficult for them to acquire rare grammatical phenomena that are scarcely attested in the input data. However, if language models nevertheless succeed in learning such rare grammatical phenomena, this, in principle, constitutes a counterargument to strong nativist claims that posit the necessity of innate biases for language acquisition. We believe that, in this sense, demonstrating partial and constrained learnability of *wanna* contraction in language models has something to offer to the study of language acquisition.

Noh et al. (2024) investigated whether standard BERT and RoBERTa models can capture the grammatical constraints on *wanna* contraction. Their results demonstrated that these models were indeed sensitive to ungrammatical instances, assigning higher surprisal values to them compared to grammatical ones. While the study confirmed that the constraint is slightly learnable by standard language models, it had two main limitations. First, Noh et al. (2024) used BERT and RoBERTa, which are not comparable to human learners in terms of the amount of linguistic input they receive. Second, although the test sentences contrasted grammatical and ungrammatical *wanna* contractions, they were constructed without incorporating frequency information from corpora that reflect actual usage of *wanna* contraction (e.g., CHILDES). Thus, this study addressed these limitations by using BabyLMs as the subjects and the CHILDES corpus as the source of the test data.

While the fundamental differences between language models and human learners should not be overlooked, models without significant advantages over human learners may offer insights into the learnability of certain language-specific biases traditionally presumed to be innate (Warstadt and Bowman, 2022). Therefore, we worked under two key assumptions. First, language models should not be provided with significant advantages in terms of input size. Second, we assume that language models lack innate linguistic biases such as those proposed by Universal Grammar. Based on these assumptions, we examined the learnability of *wanna* contraction using BabyLMs, which are not endowed with innate linguistic knowledge but are exposed to a quantity of input roughly comparable to that of human learners.

As the results show, BabyLMs generally exhibited limited sensitivity to the grammatical constraint on *wanna* contraction compared with the standard models, which were trained on vastly more data. These results indicate that learnability is partial and fragile, and may be highly dependent on input frequency and model architecture. In particular, most standard language models and several BabyLMs exhibited sensitivity to ungrammatical instances of *wanna* contractions. In addition, the fact that more models tended to show sensitivity when tested on a dataset with high-frequency embedded verbs suggests that both frequency effects and lexical exposure may have influenced their performance, as observed in children by Getz (2019). The 24 BabyLMs tested in this study were trained exclusively on textual data. Specifically, while these models received only unimodal input in the form of text, human learners were exposed to multimodal input, including non-textual information and interactions with peers and adults. Although the role of multimodal input in acquiring *wanna* contractions lies beyond the scope of this study, BabyLMs appear to be at a disadvantage compared with human learners in this respect.

While caution is warranted when generalizing from language models to humans, one way to assess the necessity of an innate bias is to test whether a model lacking that bias can still process a phenomenon hypothesized to depend on it (Warstadt and Bowman, 2022). From this perspective, language models are useful tools for probing the extent to which exposure to input alone supports the acquisition of grammatical constraints. Consequently, our findings show that the constraint on *wanna* contraction is learnable, in principle, through exposure to linguistic input, without necessarily invoking language-specific innate biases.

6 Conclusion

This study provides evidence for partial and conditional sensitivity to *wanna* contraction constraint in BabyLMs trained under developmentally plausible conditions. Specifically, we examined 24 BabyLMs from the 2024 BabyLM Leaderboard and 4 standard language models as baselines. The test sentences were divided into high- and low-frequency datasets, each containing 100 pairs of grammatical and ungrammatical *wanna* instances. The results showed that while the baseline models performed near-perfectly, the BabyLMs achieved only partial success. Models from the Strict Track, trained on larger datasets, exhibited greater sensitivity to grammaticality than those from the Strict-Small Track. Sensitivity was also higher for high-frequency input, indicating a frequency effect.

Given that the BabyLM Challenge was only launched in 2023 and the present study was carried out in 2025, certain limitations are inherent in a research framework still in its early stages of development. Situating our findings within this emerging landscape, we highlight both the challenges and the opportunities that this novel line of inquiry entails. We believe that subsequent research will not only continue to probe the potential of language models as cognitive models, but also refine the methodologies through which such evaluations are conducted. In doing so, this line of research may yield insights valuable not only for computational linguistics, but also for theoretical and experimental work on language acquisition. We hope that this study represents a meaningful step toward advancing our understanding of what language models can reveal about the learnability of grammatical knowledge.

Limitations

With respect to the language models tested, further improvement is needed. Although we prioritized cognitive plausibility by evaluating BabyLMs and constructing test sentences based on occurrences in CHILDES data, this approach alone is not sufficient to fully capture cognitively realistic learning conditions. From the perspective of language acquisition and developmental research, achieving true cognitive plausibility requires more than simply reducing the amount of training data. As language acquisition occurs through sustained interaction with the surrounding environment, experimental designs should be more sophisticated and better aligned with realistic learning conditions.

Regarding the methodology, surprisal-based evaluation in this study focuses on the prediction or masking of the sentence-final token, a design choice that may advantage encoder-based models with access to bidirectional context. Although this setup is motivated by prior psycholinguistic work, it may underestimate the sensitivity of decoder-only models to syntactic constraints.

Regarding the target phenomenon, there have been counterarguments to the traditional syntactic analysis of *wanna* contraction, especially theoretical proposals that either question or reinterpret the traditional *wh*-trace account (Postal and Pullum, 1978, 1982; Goodall, 2021). Some have also advocated for a construction-based approach that incorporates phonetic and pragmatic factors (Boas, 2004). While our test sentences were constructed using the traditional *wh*-trace account, following the widely accepted trend in prior analyses, we do not exclude the possibility that future studies may adopt alternative theoretical frameworks. Lastly, the phonological dimension of *wanna* contraction falls outside the scope of the present study, and future research incorporating speech-based models or phonetically transcribed corpora may offer a more complete picture of its learnability.

Ethics Statement

This study used only publicly available datasets and language models, with no personally identifiable information or newly collected human data. Therefore, institutional ethical approval was not required. We acknowledge that language models may reflect biases present in their training data, though the current study does not involve socially sensitive or harmful content.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments and suggestions.

References

Ben Ambridge and Lucy Blything. 2024. Large language models are better than theoretical linguists at theoretical linguistics. *Theoretical Linguistics*, 50(1–2):33–48.

Rico Behr. 2024. ELC-parserBERT: Low-resource language modeling utilizing a parser network with ELC-BERT. In *Proceedings of the 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 140–146, Miami, Florida. Association for Computational Linguistics.

György Berend. 2024. Integrating quasi-symbolic conceptual knowledge into language model pre-training. In *Proceedings of the 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 159–165, Miami, Florida. Association for Computational Linguistics.

Itamar A. Blank. 2023. What are large language models supposed to model? *Trends in Cognitive Sciences*, 27(11):987–989.

Hans C. Boas. 2004. You wanna consider a constructional approach towards wanna-contraction. In Michel Achard and Suzanne Kemmer, editors, *Language, Culture, and Mind*, pages 479–491. CSLI Publications, Stanford, CA.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Bjarne Bunzeck, David Duran, Lukas Schade, and Sina Zarrieß. 2025. Small language models also work with small vocabularies: Probing the linguistic abilities of grapheme- and phoneme-based baby llamas. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6039–6048, Abu Dhabi, United Arab Emirates. International Committee on Computational Linguistics.

Noam Chomsky. 1977. On Wh-movement. In Peter W. Culicover, Thomas Wasow, and Adrian Akmajian, editors, *Formal Syntax*, pages 91–132. Academic Press, New York.

Noam Chomsky. 1980. *Rules and Representations*. Columbia University Press, New York.

Noam Chomsky and Howard Lasnik. 1977. Filters and control. *Linguistic Inquiry*, 8(3):425–504.

Noam Chomsky, Ian Roberts, and Jeffrey Watumull. 2023. The false promise of ChatGPT. *The New York Times*. Opinion.

Stephen Crain and Rosalind Thornton. 1998. *Investigations in Universal Grammar: A Guide to Experiments on the Acquisition of Syntax and Semantics*. MIT Press, Cambridge, MA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lukas Edman. 2024. Babylm 2024. `https://huggingface.co/collections/leukas/babylm-2024-66e1bae0dc9eb134560be8ff`. Accessed: 2025-09-14.

Danny Fox and Roni Katzir. 2024. Large language models and theoretical linguistics. *Theoretical Linguistics*, 50(1–2):71–76.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Heidi R. Getz. 2019. Acquiring Wanna: Beyond universal grammar. *Language Acquisition*, 26(2):119–143.

Geoffrey Goodall. 2021. Wanna-contraction as restructuring. In Grant Goodall, editor, *Theory and Experiment in Syntax*, pages 53–68. Routledge, London.

Zachary Goriely, Ricardo D. Martinez, Andrew Caines, Lisa Beinborn, and Paul Buttery. 2024. From babble to words: Pre-training language models on continuous streams of phonemes. *arXiv preprint arXiv:2410.22906*.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 159–166, Pittsburgh, Pennsylvania. Association for Computational Linguistics.

John Hale. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9):397–412.

Michael Y. Hu, Aaron Mueller, Charles Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan G. Wilcox. 2024. Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. *arXiv preprint arXiv:2412.05149*.

Hyejin Hwang. 2023. Wanna contraction in first language acquisition, child second language acquisition, and adult second language acquisition. *Bilingualism: Language and Cognition*, 27(3):322–333.

Yoshifumi Ito. 2018. Acquisition of contraction constraint by japanese learners of english. *Journal of the Pan-Pacific Association of Applied Linguistics*, 22(1):19–41.

Anna A. Ivanova, Abhishek Sathe, Benjamin Lipkin, Utkarsh Kumar, Samira Radkani, Taylor H. Clark, Cameron Kauf, Jiahui Hu, R. T. Pramod, Gabriel Grand, Viktor Paulun, Maria Ryskina, Ekin Akyürek, Ethan Wilcox, Nadia Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Joshua B. Tenenbaum, and Jacob Andreas. 2024. Elements of world knowledge (EWOK): A cognition-inspired framework for evaluating basic world knowledge in language models. *arXiv preprint arXiv:2405.09605*.

Pablo Contreras Kallens, R. D. Kristensen-McLachlan, and Morten H. Christiansen. 2023. Large language models demonstrate the potential of statistical learning in language. *Cognitive Science*, 47(3):e13256.

Roni Katzir. 2023. Why large language models are poor theories of human linguistic cognition: A reply to piantadosi. *Biolinguistics*, 17:1–12.

George Lakoff. 1970. Global rules. *Language*, 46(3):627–639.

David Lightfoot. 1976. Trace theory and twice-moved NPs. *Linguistic Inquiry*, 7(4):559–582.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Erik Lucas, Daniel Gaines, Tejaswi Rao Kosireddy, Kun Li, and Timothy C. Havens. 2024. Using curriculum masking based on child language development to train a large language model with limited training data. In *Proceedings of the 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 221–228, Miami, Florida. Association for Computational Linguistics.

Qihang Niu, Jiarui Liu, Zongyan Bi, Pei Feng, Baolin Peng, Keqi Chen, Ming Li, L. K. Q. Yan, Yuhui Zhang, Chunhui Yin, Chenyu Fei, Tian Wang, Yichi Wang, Shuaichen Chen, and Minjie Liu. 2024. Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges. *arXiv preprint arXiv:2409.02387*.

Kangsan Noh, Eunjeong Oh, and Sanghoun Song. 2024. Testing language models' syntactic sensitivity to grammatical constraints: A case study of wanna contraction. *Frontiers in Communication*, 9:1442093.

Steven T. Piantadosi. 2024. Modern language models refute chomsky's approach to language. In Edward Gibson and Moshe Poliak, editors, *From Fieldwork to Linguistic Theory: A Tribute to Dan Everett*, volume 15 of *Empirically Oriented Theoretical Morphology and Syntax*, pages 353–414. Language Science Press, Berlin.

Paul M. Postal and Geoffrey K. Pullum. 1978. Traces and the description of english complementizer contraction. *Linguistic Inquiry*, 9(1):1–29.

Paul M. Postal and Geoffrey K. Pullum. 1982. The contraction debate. *Linguistic Inquiry*, 13(1):122–138.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Joel T. Rotenberg. 1978. *The Syntax of Phonology*. Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA.

David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. 2023. Trained on 100 million words and still in shape: BERT meets the british national corpus. *arXiv preprint arXiv:2303.09859*.

Jean-Luc Tastet and Ildar Timiryasov. 2024. BabyLLaMA-2: Ensemble-distilled models consistently outperform teachers with limited data. *arXiv preprint arXiv:2409.17312*.

Nikolaos Theodoropoulos, George Filandrianos, Vassilis Lyberatos, Maria Lymperaiou, and Gerasimos Stamou. 2024. BERTtime stories: Investigating the role of synthetic story data in language pre-training. *arXiv preprint arXiv:2410.15365*.

Ildar Timiryasov and Jean-Luc Tastet. 2023. Baby LLaMA: Knowledge distillation from an ensemble of teachers trained on a small dataset with no performance penalty. *arXiv preprint arXiv:2308.02019*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5998–6008.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Alex Warstadt and Samuel R. Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In Uriel Cohen Priva, Isabelle Charnavel, and Brian Dillon, editors, *Algebraic Structures in Natural Language*, pages 17–60. CRC Press, Boca Raton, FL.

Alex Warstadt, Leshem Choshen, Aaron Mueller, Ethan Wilcox, Adina Williams, and Chengxu Zhuang. 2023a. Call for papers – the BabyLM challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv preprint arXiv:2301.11796*.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhavan V. Paranjape, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023b. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

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 Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2022. Using computational models to test syntactic learnability. *Linguistic Inquiry*, 55(4):805–848.

Xingwei Yu, Bin Guo, Shimin Luo, Jun Wang, Tongtong Ji, and Yuxuan Wu. 2024. AntLM: Bridging causal and masked language models. *arXiv preprint arXiv:2412.03275*.

Andrea Zukowski and J. Larsen. 2011. Wanna contraction in children: Retesting and revising the developmental facts. *Language Acquisition*, 18(4):211–241.