Can Language Models Induce Grammatical Knowledge from Indirect Evidence?

Miyu Oba[1] **Yohei Oseki**[2] **Akiyo Fukatsu**[2] **Akari Haga**[1]
Hiroki Ouchi[1] **Taro Watanabe**[1] **Saku Sugawara**[3]
[1]Nara Institute of Science and Technology
[2]The University of Tokyo [3]National Institute of Informatics
{oba.miyu.ol2,haga.akari.ha0,hiroki.ouchi,taro}@is.naist.jp
{oseki,akiyofukatsu}@g.ecc.u-tokyo.ac.jp
saku@nii.ac.jp

Abstract

What kinds of and how much data is necessary for language models to induce grammatical knowledge to judge sentence acceptability? Recent language models still have much room for improvement in their data efficiency compared to humans. This paper investigates whether language models efficiently use indirect data (*indirect evidence*), from which they infer sentence acceptability. In contrast, humans use indirect evidence efficiently, which is considered one of the inductive biases contributing to efficient language acquisition. To explore this question, we introduce the Wug Indirect Evidence Test (WIDET), a dataset consisting of training instances inserted into the pre-training data and evaluation instances. We inject synthetic instances with newly coined *wug* words into pretraining data and explore the model's behavior on evaluation data that assesses grammatical acceptability regarding those words. We prepare the injected instances by varying their levels of indirectness and quantity. Our experiments surprisingly show that language models do not induce grammatical knowledge even after repeated exposure to instances with the same structure but differing only in lexical items from evaluation instances in certain language phenomena. Our findings suggest a potential direction for future research: developing models that use latent indirect evidence to induce grammatical knowledge.

1 Introduction

Recent advances in language models, such as those from the GPT and Llama families (OpenAI, 2024; Meta, 2024), have shown remarkable progress in various tasks. These models are trained on extremely large datasets, on a scale thousands of times greater than the amount of data children are exposed to in developing grammatical knowledge comparable to that of adults (Warstadt et al., 2023). This suggests substantial potential for improving their learning efficiency.

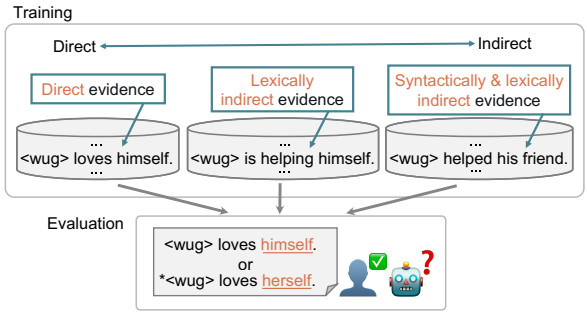

Figure 1: The indirectness of evidence. Direct evidence refers to instances identical to previously observed ones. Lexically indirect evidence targets the same linguistic knowledge but differs in lexical items. Syntactically & lexically indirect evidence is different in both their syntactical and lexical items.

According to Pearl and Mis (2016), humans acquire language using *indirect* evidence, in addition to *direct* evidence, which is considered one of the inductive biases contributing to efficient language acquisition. As illustrated on the left side of Figure 1, when humans encounter the sentence "<wug> loves himself.", they can correctly judge the grammatical acceptability between "<wug> loves himself." and "*<wug> loves herself." Such observed sentences are referred to as *direct* evidence. Conversely, in the middle and right sides of the figure, we assume that humans are not exposed to such direct evidence. However, if they observe sentences from which they can make some inference for a correct judgment, such sentences are called *indirect* evidence. For example, humans might hypothesize that "him(self)" in the sentence "<wug> is helping himself." refers to <wug>, or that the pronoun "his" in "<wug> helped his friend." indicates <wug> has a masculine property.

However, it remains still unclear how the degree of indirectness in observed instances affects the number of occurrences required for language models to induce grammatical knowledge. Pre-

vious work has investigated how language models learn grammatical knowledge based on the appearance of items in training data focusing on the word frequency effect (Wei et al., 2021; Yu et al., 2020) or generalization to unseen instances (Patil et al., 2024; Misra and Mahowald, 2024; Leong and Linzen, 2024) through few-shot learning or pretraining on corpora filtered by specific linguistic constructions. However, those methods face a limitation in identifying ways to enhance the model's learning efficiency.

In this work, we explore the degree of indirectness and the amount of data needed for language models to induce linguistic generalization. To address this question, we introduce the Wug InDirect Evidence Test (WIDET), a dataset containing additional indirect training and evaluation instances. We train language models on pretraining data incorporating the indirect training instances. We then evaluate their linguistic generalization across seven different phenomena, including anaphor agreement, transitivity, and subject-verb agreement. These phenomena require language models to comprehend diverse properties and multiple parts of speech of specific words to judge their acceptability. To control the number of observed indirect training instances, we inject synthetic instances with newly coined words into pretraining data. Following Berko (1958), we refer to these words that do not appear in the original vocabulary and data as *wug* words.[1] We use various synthetic data as additional indirect training instances, each differing in the degree of lexical and syntactic indirectness as well as the number of observations.

We find that language models generalize linguistic knowledge from training instances identical to correct evaluation instances, though their data efficiency varies across different linguistic phenomena. This variation is likely influenced by the number of words between the *wug* and the words that act as cues for the model to learn its properties. We surprisingly observe that the language models do not induce grammatical knowledge in certain phenomena, even in instances that only differ in lexical items. Syntactically indirect instances rarely induce the model's generalization.

Given that the distances between the *wug* and the cue words to learn its properties might cause inefficiency in the models' learning, we conduct a detailed analysis of indirect instances with complex interference, using anaphor gender agreement as a case study. We examine whether these instances affect the generalization, considering three factors related to attractors and distance, finding that when the language models are trained on the instances with complex interference, they hit a plateau in learning after sufficient observations.

Those findings from our controlled and comprehensive experiments suggest that, at least in our small-scale settings, language models do not generalize in a human-like manner even from the data with a degree of indirectness that seems intuitively manageable for humans, depending on language phenomena. Our work contributes insights into language models' capacity to use indirect evidence for learning. To advance this in future research direction: Implement a model that can use indirect evidence, enabling data-efficient language acquisition comparable to that of humans.[2]

2 Background

2.1 Evidence in Language Acquisition

In the field of language acquisition, the information used to learn grammatical knowledge is referred to as *evidence*. Positive (negative) evidence refers to information in data that indicates what is acceptable (unacceptable) in a language, and it has been argued that humans rely solely on positive evidence to acquire their language (Chomsky, 1993). Pearl and Mis (2016) further distinguishes indirect positive evidence from direct positive evidence. Direct positive evidence indicates the information present in the data observed by the learner and used for learning, with the assumption that its usage by speakers guarantees grammaticality (the left side of Figure 1). Indirect positive evidence, by contrast, refers to information that requires a learner to infer what is grammatical in the language from observed data (the middle and right side of Figure 1). They argue that, in addition to direct positive evidence, indirect positive evidence potentially plays a significant role in efficient language acquisition. While the previous literature explores humans' capacity, it is still unclear whether language models induce linguistic generalization from such evidence.

[1] The original *wug* used in Berko (1958)'s work is not exactly the same as our setting to create controlled instances. Details are discussed in Section 7.1.

[2] WIDET is publicly available at https://github.com/nii-cl/widet.

2.2 Analysis of Language Models in Learning Grammatical Knowledge

Previous studies have focused on how language models learn grammatical knowledge based on the appearance of target lexical items in training data. Yu et al. (2020) evaluate models' performance on grammatical tasks using minimal pairs including specific target words and few-shot learning on sentences including unseen words. Wei et al. (2021) train models on data where the frequency of instances including specific verbs is manipulated to evaluate their generalization to verb inflections. Recent studies have focused on indirect evidence (Misra and Mahowald, 2024; Leong and Linzen, 2024; Patil et al., 2024), exploring the availability of indirect evidence in language models by training them from scratch on filtered data. These data include specific distinctive linguistic phenomena, such as AANN construction (Misra and Mahowald, 2024) and passivization (Leong and Linzen, 2024), and systematic phenomena from BLiMP (Warstadt et al., 2020b).

3 Motivations

3.1 Experiment Design

While the previous studies in Section 2.2 each offer valuable insights into how language models generalize to unseen instances from various perspectives, our goal in this work is to explore the impact of the degree of indirectness on data efficiency, with the aim of identifying ways to enhance the model's learning efficiency. Specifically, we examine how the number of instances required for language models to induce grammatical knowledge changes as the degree of indirectness in the training instances increases. To achieve this, we assume that experiments have to meet the following requirements:

Various Degrees of Indirectness in a Single Linguistics Phenomenon To investigate the impact of the degree of indirectness on the number of observations needed for grammar acquisition, we employ two graduated types of indirectness, lexical and syntactic, in addition to direct evidence. Most prior research focuses on a single degree of indirectness for a given linguistic phenomenon.

Various Number of Observations Given our aim for data efficiency, we need to quantify how much the required amount of data for language models to induce grammatical knowledge increases due to indirectness. We employ six different observation counts, ranging from 0 to 100. Previous studies focusing on indirect evidence are limited in their ability to quantify changes in the number of observations required, as they do not take into account the frequency effect.

Various Linguistics Phenomena We explore whether the two aspects mentioned above occur universally across linguistic phenomena or are specific to certain phenomena. We employ seven types of linguistic phenomena, each with target words consisting of several different parts of speech. Most of the previous work, except for Patil et al. (2024), focuses on one or two phenomena.

Inserting Sentences Containing Words that do not Appear in Pretraining Data Considering phenomena like anaphor gender agreement, to judge the acceptability of a sentence, language models are expected to understand the gender properties of the antecedent (*target word*) of the reflexive. To count the number of observations for language models to induce grammatical knowledge, we need to concisely count how many times the language models encounter a sentence containing the target word before they understand the properties of the word. For conventional approaches to ablate certain lexical items existing in corpora, the (sub)word of the target word may appear in the sentence other than the removed one, making it difficult to count the observations accurately. To precisely control the number of observations of the target word, we employ sentences containing target words that have not appeared in the corpus.

3.2 Inserting Instances with Newly Coined Words

We employ newly coined words (*wugs*) to introduce additional instances including words that do not appear in pretraining data. The advantages include:

- Handling the occurrences of target lexical items may not eliminate their influence from the pretraining corpus. To fully negate the effect of a lexical item, all variants sharing the same stem or subword would need to be removed, which is complex and risks significantly distorting the natural corpus distribution.

- When automatically generating *wugs*, we can adequately control their frequency and evidence strength, including their tokenization. Since our

Phenomenon	*Evd*	Training instance	Evaluation instance
Anaphor gender agreement (ANA.GEN.AGR)	DE	<wug#n> has devoted herself	<wug#n> has devoted herself
	LexIE	<wug#n> is painting herself	*<wug#n> has devoted himself
	SynIE	<wug#n> judges her work	
Anaphor number agreement (ANA.NUM.AGR)	DE	the <wug#n> didn't see themselves	the <wug#n> didn't see themselves
	LexIE	the <wug#n> can reward themselves	*the <wug#n> didn't see itself
	SynIE	the <wug#n> loved their toy	
Transitive (TRANS.)	DE	some trees <wug#n>ed the car	some trees <wug#n>ed the car
	LexIE	no street can <wug#n> the city	*some trees <wug#n>ed
	SynIE	every lion hunts what no prey can <wug#n>	
Intransitive (INTRANS.)	DE	many rivers should <wug#n>	many rivers should <wug#n>
	LexIE	each ethic might <wug#n>	*many rivers should <wug#n> dogs
	SynIE	a man corrects that the answer will not <wug#n>	
Determiner-Noun agreement (D-N AGR)	DE	the senators use this <wug#n>	the senators use this <wug#n>
	LexIE	a window will open this <wug#n>	*the senators use these <wug#n>
	SynIE	the <wug#n> sells the house	
Subject-Verb agreement (V) (S-V AGR (V))	DE	the <wug#n> are leaving any traces	the <wug#n> are leaving any traces
	LexIE	the <wug#n> climb few ladders	*the <wug#n> is leaving any traces
	SynIE	each key can open those <wug#n>	
Subject-Verb agreement (S) (S-V AGR (S))	DE	the book <wug#n> a shelf	the book <wug#n> a shelf
	LexIE	every chocolate <wug#n> several bars	*the books <wug#n> a shelf
	SynIE	the deer that trails the head <wug#n> a herd	

Table 1: Linguistic phenomena and instances. The sentences starting with * are ungrammatical.

Phenomenon	POS	Gen.	Num.	(In)Transitive	Long agr
ANA.GEN.AGR.	noun	✓	–	–	✓
ANA.NUM.AGR	noun	–	✓	–	✓
TRANS.	verb	–	–	✓	–
INTRANS.	verb	–	–	✓	–
D-N AGR	adj	–	✓	–	–
S-V AGR (V)	verb	–	✓	–	–
S-V AGR (S)	noun	–	✓	–	–

Table 2: Properties to judge evaluation data. POS denotes part-of-speech. Gen./Num. denotes gender/number. Long agr. is whether a long agreement is required.

aim here is to control the minimal information observable by the model, synthetic data allows for the elimination of noises.

- Our approach is a form of data augmentation, that does not require any modification of lexical items or sentences in the corpora. Hence, it can be easily applied to other corpora and models.

While using artificial languages in analyzing language models is tackled by previous work (White and Cotterell, 2021; Ri and Tsuruoka, 2022), our approach is different in that we use artificial instances only at the token level by introducing a word *wug* to insert them into a natural corpus.

4 Wug InDirect Evidence Test (WIDET)

This section describes how we construct additional training and evaluation instances, which comprise our dataset, WIDET. Following targeted syntactic evaluation (Linzen et al., 2016; Marvin and Linzen, 2018; Warstadt et al., 2020b), we employ minimal pair paradigm where pairs of sentences minimally differ in target words. The examples of instances are listed in Table 1.

4.1 Linguistic Phenomena

We employ the seven different linguistic phenomena listed in Table 1, which we selected from the benchmark BLiMP (Warstadt et al., 2020b)[3]. As shown in Table 2, the phenomena vary in their properties, so that we can analyze models' behavior from diverse perspectives. Since our selection criteria are based on whether understanding the properties of a single word is sufficient to judge the linguistic phenomena correctly, we can only cover limited linguistic phenomena. We anticipate phenomena related to island effects, for instance, to be beyond this scope.

4.2 Newly Coined Word *Wug*

We employ the tag <wug#n> as a newly coined word to conduct controlled experiments using words that never appeared in the pretraining corpus. This approach does not entirely align with the policy in Berko (1958), which employed words like *wug* and *wuz* that are newly coined but phono-

[3]Appendix A.1 details the specific phenomena referenced from BLiMP in this work.

logically natural in the target language by using actual subwords. One concerning issue with Berko (1958)'s policy is that the actual subwords can provide models with clues for correct grammatical judgment, for example, by their occurrence in specific positions. While using actual subwords could help models access grammatical knowledge needed for accurate judgment, it complicates evaluating the models' true ability to learn from indirect evidence. To avoid its possible effects, we instead use the artificial tag <wug#n>. We analyze the differences between the conditions using the tag and the original *wug* in Section 7.1.

4.3 Indirectness of Additional Training Instances

We define the following three degrees of indirectness (DE, LexIE, and SynIE). The difficulty increases in the order of DE, LexIE, and SynIE:

Direct Evidence (DE) An instance identical to the correct evaluation instances. We assume that the properties of *wug* in an evaluation instance are learned by referencing the training instance that shares the same syntactical and lexical items as the evaluation instance.

Lexically Indirect Evidence (LexIE) An instance that conveys the same syntactic structure as the evaluation instance but uses different lexical items. We assume that the properties of *wug* in an evaluation instance are learned by referencing training instances with the same usage but different lexical items from those in the evaluation instance.

Syntactically Indirect Evidence (SynIE) An instance that reveals the target linguistic feature with different syntactic and lexical items from evaluation instances. The properties of *wug* in an evaluation instance are learned by referencing the training instance with different syntactic and lexical items from those in the evaluation instance.

4.4 Training and Evaluation Template

We prepare 200 template pairs for each linguistic phenomenon. Each template has three different sets of tags, resulting in $200 \times 3 = 600$ pairs.

We anticipate that quantifiers and determiners can influence linguistic generalization, making it unclear whether language models rely on the properties of verbs and reflexive pronouns, quantifiers, and determiners, or other factors as clues for judgment, while previous studies have paid limited attention to this (Patil et al., 2024). To mitigate such effects, for number agreement, we added <wug#n> without any suffixes to these sentences, expecting the models to infer that <wug#n> is an inflected form based on the sentence structure in which they are embedded. We explore their effects in the model's generalization in Section 7.1. For the noun subject of S-V AGR (V) and ANA.NUM.AGR, we avoid any quantifiers and determiners other than "the". Due to the same reason, for the verb in S-V AGR (S), we only employ the present tense and do not employ any auxiliary verbs and tense suffixes. We ensured that <wug#n> was used the same word (i.e., the tag with the same id) in a pair, both grammatical and ungrammatical sentences because we want the same occurrence of the *wug* in the training data.

4.5 Data Generation with LLM

To create varied degrees of and balanced corpus, we use GPT-4 Turbo in OpenAI API to generate the training and evaluation templates. To generate balanced training instances with different properties, we generate them separately based on concerning properties, (e.g., feminine and masculine pronouns have the same percentage in ANA.GEN.AGR.). We prompt the GPT-4 to generate balanced, diverse, and deduplicated sentences. We generate evaluation instances and training instances for indirect evidence (LexIE, SynIE) with three different prompts. Subsequently, we get DE by extracting the correct sentence in generated evaluation instances. We generate the sentences with placeholders [WUG] and we replace [WUG] with the tag <wug#n>, where the index number n distinguishes the coined words (e.g., <wug#124>). The example of prompts and detailed procedures are shown in Appendix A.4.

5 Experiments and Results

5.1 Settings

Pretraining Data We randomly sample 675k sentences (16M words) from English Wikipedia articles and use them as pretraining data.[4] We inject the additional training instances into the data. The detailed preprocessing steps and additionally injected training instances are described in Appendix A. We shuffle and deduplicate sentences and remove ones containing fewer than two words. The

[4]Retrieved from https://github.com/phueb/BabyBERTa.

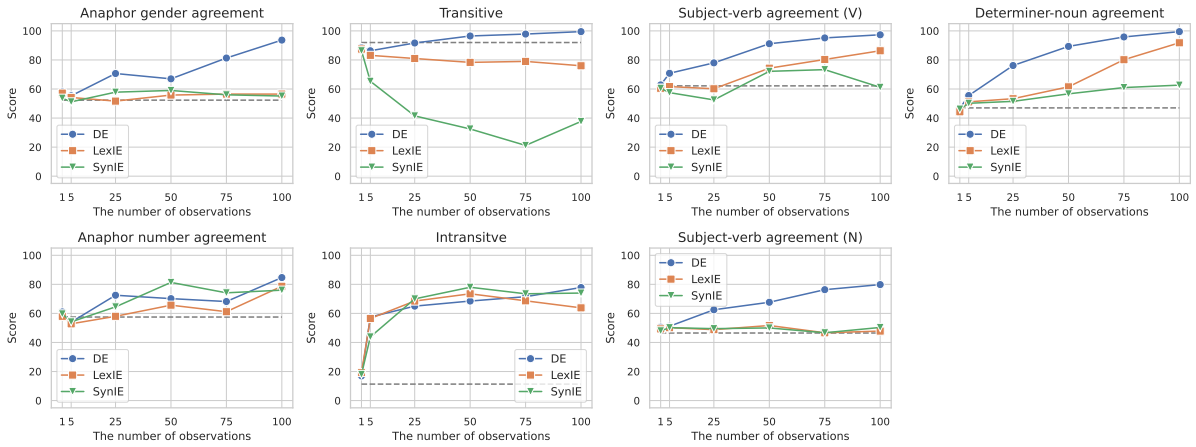

Figure 2: The results (accuracy; %) of experiments for language phenomena and evidence. The gray dot lines indicate the model's scores trained on pretraining data without any additional instances (n=0).

data is then lowercase, and periods are removed from the sentences.

Frequency of Additional Instances We compare the language models trained on the pretraining data injected indirect instances that appear n times ($n = 0, 1, 5, 25, 50, 75, 100$) for each instance.

Models We use BabyBERTa (Huebner et al., 2021), which is a minimal variant of RoBERTa (Liu et al., 2019). We modify some hyperparameters due to the pretraining data size. More detailed information is shown in Table 6. We train the tokenizer from scratch on the pretraining data, adding the tags to the vocabulary so that the tokenizer treats each tag as one token.

Evaluation Metrics We use the accuracy of selecting the correct sentence as our evaluation metric. We employ pseudo-likelihood (Salazar et al., 2020)[5] normalized by token length because we use evaluation sentences containing the sentence pair each of which has different token lengths. [6]

5.2 Results

We review the main results by answering our research questions: (i) What degree of and how much data do language models need to acquire grammatical knowledge to judge the acceptability of a sentence? (ii) Are observations showing similar trends in broader categories of linguistic phenomena? The results are shown in Figure 2.

[5]We use the source code in https://github.com/babylm/evaluation-pipeline-2023.

[6]Normalization by token length may still result in token-biases (Ueda et al., 2024).

Direct Evidence As for DE, increasing the number of observations generally contributed to linguistic generalization in language models. However, the extent of improvement varied across different linguistic phenomena. In ANA.GEN.AGR and ANA.NUM.AGR, the score increased more gradually, particularly between 25 and 75 occurrences, compared to the other agreement phenomena. This difference might be due to anaphor agreement, which often involves a longer distance between the target words and the words with properties necessary for correct judgment. We thoroughly examine the effects of distance and attractors in Section 6.

Lexically Indirect Evidence In about a half of the phenomena, D-N AGR, S-V AGR (V), ANA.NUM.AGR, and INTRANSITIVE, LexIE induces generalization more slowly but steadily than DE. However, in the remaining half of the phenomena, the language models do not acquire the grammatical knowledge necessary to correctly judge acceptability. This result is surprising because LexIE differs only in lexical items from a correct sentence in the evaluation and shares the same syntactical structure. This trend cannot be explained by the properties of Table 2.

Syntactically Indirect Evidence In most phenomena, the models fail to induce SynIE generalization; the increase in the number of observations did not improve generalization but merely extended learning time. In TRANSITIVE, the accuracy of SynIE drastically decreases inversely with the number of observations. This intriguing phenomenon is likely due to the heuristics of the language model. The final word in the training in-

Interf.	Evd.	Training instance
Attractor type (AT)	DE	<w> loves herself
	AT0	<w> **helping the child** loves herself
	AT1	<w> **helping the man** loves herself
	AT2	<w> **helping him** loves herself
Attractor number (AN)	DE	<w> loves herself
	AT1	<w> **helping the man** loves herself
	AN0	<w> **helping the man to see the dad** loves herself
	AN1	<w> **helping the man for the king to see the dad** loves herself
	AN2	<w> **helping the man for the son of the king to see the dad** loves herself
Distance (DT)	DE	<w> loves herself
	AT0	<w> **helping the child** loves herself
	DT0	<w> **who helps the child** loves herself
	DT1	<w> **whose cat helps the child** loves herself
	DT2	<w> **whose cat helps the child who finds the teachers** loves herself

Table 3: Interference types and training instances used in the analysis. <w> corresponds to <wug#n>.

stances (see Table 1) is the <wug#n>, whereas it is an actual direct object noun in the correct evaluation sentences. This suggests that the language model might exhibit linear generalization (Mueller et al., 2022; McCoy et al., 2020), which differs from the human-like hierarchical generalization. The model seems to judge correctness based on whether certain words follow the <wug#n>, even though the *wug* should be recognized as a transitive verb because the relative pronoun "what" is its object. This implies that instances requiring complex hierarchical inference may hinder generalization.

Overall Our findings mainly suggest that language models do not sufficiently induce linguistics generalization from indirect positive evidence, especially SynIE, while they induce it from direct evidence. Wei et al. (2021) find that their results support the Reduce Error Hypothesis (Ambridge et al., 2015), where high-frequency words are learned better. The results in our work also support the hypothesis in DE, but in LexIE and SynIE, not all linguistic phenomena support it.

6 Analysis with More Indirect Instances

In Section 5, DE induced the model's linguistic generalization but its data efficiency varies by linguistic phenomena. For anaphor agreement, the models' learning is more apt to reach a plateau in 25 – 75 observations compared to other phenomena (See the figure for anaphor agreement in Figure 2). This stagnation might be caused by the longer distance between the *wug* and the reflexives, whereas the relevant items are adjacent to each other in other phenomena such as TRANSITIVE. To corroborate this negative effect of long distance on learning, we employ more indirect agreement instances to investigate whether the long distance hinders linguistic generalization on ANA.GEN.AGR in language models.

The difficulty of long-distance agreement is caused by attractors and distance (Linzen et al., 2016). Agreement attractors indicate the intervening words that distract the learner from judging the correct agreement (Giulianelli et al., 2018). When language models judge the gender agreement, they would check if the word "<wug#n>" corresponds to the gender of the reflexive. *Distance* refers to the number of the words intervening between the antecedent "<wug#n>" and "herself". *Attractor* indicates the competing words (e.g., "man" in the case of AT1 in Table 3) that distract learners from judging the agreement.

The language models' grammatical knowledge concerning long-distance dependencies has been investigated in previous studies (Giulianelli et al., 2018; Li et al., 2023), and these studies argue that the models can indeed acquire the knowledge of long-distance agreement. However, the overall results on anaphor agreement in this study suggest that further investigation is required to reveal the relationship between models' performance and the distance of items relevant to correct judgment. For this purpose, we conduct a fine-grained analysis using synthetic sentences varying the distance between *wugs* and reflexive pronouns.

6.1 Target Phenomena

We compare the models trained on the corpus with additional instances of anaphor gender agreement, from the perspective of the attractor type, number, and distance as below. Table 3 lists all kinds of training instances compared in this analysis.

To create the instances, we use GPT-4 to generate nouns differing in gender and number and sample the designated number of items from these generated items. For feminine and masculine nouns, we collect 100 nouns each. From the generated items, we first select 25 nouns for each gender. Then, we create both the singular and plural forms

of the selected words and double them to create minimal pairs. The prompt is shown in Appendix A.4. Additionally, we also collect 100 neutral nouns such as teacher and child. The verb that we newly employ is collected from LexIE in ANA.GEN.AGR to avoid duplication.

Attractor Type (AT) We investigate whether attractors downgrade the linguistic generalization in ANA.GEN.AGR and how their distract strength affects the models' acquisition of anaphor agreement. DE indicates the direct instances examined in Section 5, which does not have any attractors and works as a baseline here. AT0 includes neutral common nouns, while AT1 employs common opposite-gender nouns, and AT2 uses opposite-gender pronouns. We assume that the magnitude of attractors' interference follows the order AT0 < AT1 < AT2, given that the more similar their properties are to reflexives, the more distracting they will be.

Attractor Number (AN) We examine whether the number of attractors affects the model's acquisition. We use the gender common nouns as attractors. DE works as a baseline because it has no attractors. We expect that the more attractors there are, the more difficult it is to generalize correctly.

Distance (DT) We analyze the effect of distance on the model's acquisition. We assume that the more distance intervening between *wug* and reflexive, the more difficult it is to judge sentence acceptability. We use neutral nouns there to explore the effect of the number of words genuinely.

6.2 Results

As shown in Figure 3, After 100 observations in all viewpoints, SynIE, with the shortest distance and no attractors, got the highest scores, while in midway observations this tendency does not happen. The most difficult instances in each interference lead to the language model's lowest score, after their 100 observations. AT2, including an opposed pronoun as an attractor, particularly shows unstable generalization. We initially expected that instances with longer distances and more attractors would interfere more strongly with the models' generalization. However, this tendency was not observed in the experiment. To the question of whether the instances with long-distance agreement induce linguistic generalization, these results answer that with the larger number of observations, the model's generalization relatively hits a plateau.

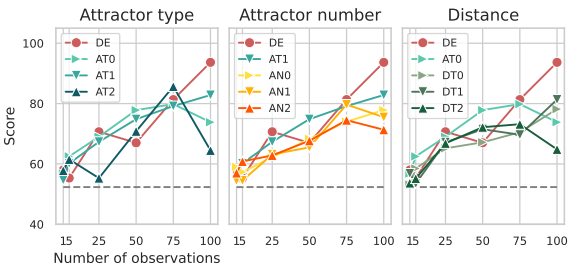

Figure 3: Models' scores for more indirect instances.

7 Discussion

7.1 Considering *Wug* Creation

In this work, we use newly coined words that do not appear in the original vocabulary, following Berko (1958). Still, our used *wug* has some gap from the original one. In the original *wug* test, they use the words that do not exist in the language but conform to the phonological rule in the language. In contrast, we use the tag <wug#n> as *wug* in those experiments. Since the original *wug* is more phonologically natural, and the subwords are in the existing vocabulary, the original setting is closer to the environment of human language acquisition. On the other hand, to conduct controlled experiments on the number of instances that the model observed, the setting might not be suitable because this is far from the settings where a certain word is never encountered. We used the tag <wug#n>. In this section, we compare our method (*tag* method) and the original method (*wug* method) to explore the difference in their impact on the model's linguistic generalization.

***Wug* Generation** We create *wug* using pseudoword generator Wuggy.[7] and choose 1.2k nouns from sample data taken from the one billion-word Corpus of Contemporary American English (COCA).[8] To create *wug*-like words, we use the nouns to output four pseudo words for one noun and randomly select one pseudo noun. We prepare $200 \times 3 = 600$ pseudo words, each 200 of which are used separately (*wug_v1–wug_v3*) because we expect that different *wug*s have different subwords and they can show different results. [9] We use those

[7] https://github.com/WuggyCode/wuggy.

[8] Downloaded from https://www.wordfrequency.info/samples/words_219k.txt.

[9] On the other hand, for *tag* and *tag w/ morph.*, we show the results of only one model, because the different *tag*s <wug#n> have the same parameters and they actually show the same results.

N	*wug* method	Phenomenon		
		ANA. NUM. AGR	D-N AGR	S-V AGR (V)
0	*tag*	57.5	47.0	62.2
	tag w/ morph.	59.0	80.5	83.3
	wug_v1	81.3	89.5	86.7
	wug_v2	81.2	91.2	86.0
	wug_v3	81.5	88.7	85.0
25	*tag*	72.5	76.2	78.0
	tag w/ morph.	94.0	99.5	91.3
	wug_v1	92.3	87.7	90.2
	wug_v2	90.5	87.7	88.5
	wug_v3	90.5	87.5	86.5

Table 4: Scores calculated by the models trained on the pretraining data with indirect instances of different *wug* creation methods. *N* is the number of observations.

pseudo nouns instead of the tag in the same way as in the previous experiments.

Settings We target three phenomena, ANA.NUM.AGR, D-N AGR, and S-V AGR (V), the *wug* of which is considered as common nouns. No inflectional morphemes are added to plural common nouns in the *tag* method while the morphemes are added to plural common nouns in the *wug* method. For ablation, we prepare the tag with inflectional morphemes (*tag w/ morph.* method), which employs the tag <wug#n> same as the *tag* method but uses inflectional morphemes same as the *wug* method. We compare the models trained on the pretraining data with the *tag* method, the *wug* methods, and *tag w/ morph.* method. Other settings are the same as Section 5.

Results Table 4 shows the scores of the *tag*, *tag w/ morph.*, and three sets of *wug*. In the *wug* and *tag w/ morph.*, the language models correctly judge the acceptability of sentences, mostly more than 80–90%, surprisingly with the data that includes zero additional instances. This result is probably because language models determine whether a word is singular or plural, based on whether an inflection morpheme "s" follows it, even if the word is novel. This occurs with both novel words and novel subword combinations, but the impact is greater with the latter, comparing the two methods. In addition, despite our expectation that different subword combinations show different results, we observed no large score variances among the three vocabulary sets except for 25 times in ANA.NUM.AGR. From those results, we found a trade-off between the settings plausible for human language acquisition and strictly controlled settings. We prioritized the latter in this work, but the direction to the former is also a good setting depending on the research questions.

Phenomenon	Std	Score
ANA.GEN.AGR	0.02	51.3 ± 0.95
	0.002	55.5 ± 1.73
ANA.NUM.AGR	0.02	59.7 ± 2.44
	0.002	64.4 ± 2.84
TRANSITIVE	0.02	90.2 ± 1.57
	0.002	90.0 ± 1.15
INTRANSITIVE	0.02	12.7 ± 1.53
	0.002	12.0 ± 0.60
D-N AGR	0.02	47.4 ± 1.39
	0.002	48.9 ± 1.68
S-V AGR (V)	0.02	56.4 ± 5.23
	0.002	54.7 ± 1.78
S-V AGR (S)	0.02	49.1 ± 2.98
	0.002	49.4 ± 1.19

Table 5: Scores (mean±std) of language models with different seeds and standard deviation of the initializers.

7.2 Zero Observations of *Wug*

While a tag <wug#n> is added to the vocabulary, its parameters in language models are randomly initialized. If the language models never encounter sentences containing this tag during training, its parameters still remain in their initialized state, which may lead to varying results in language models depending on factors such as the initializer's standard deviation (std) and the random seed used. To verify this effect, we compare the language model using the default std of the initializer for all weight matrices (std $= 0.02$) to that with one-tenth std (std $= 0.002$), using three kinds of seeds. Table 5 shows that the deviation of scores is smaller in the model using one-tenth std for initializer compared to the model using the default std. This finding implies that a smaller std can enhance the stability of the results. However, an excessively small std may risk negatively affecting the training process. Hence, we employ default std in the current work.

8 Conclusion

We investigate the degree of indirectness and the amount of data required to induce human-like linguistic generalization in language models. We found that language models do not induce human-like linguistic generalization even with a degree of indirectness that seems intuitively manageable for humans, depending on language phenomena. This limitation indicates a direction for future studies: implementing a model that can use indirect evidence, which will lead to data-efficient language acquisition comparable to that of humans.

Limitations

We recognize the following limitations in this study:

Linguistic Knowledge by Function Words We generate synthetic instances only for linguistic phenomena concerning content words such as nouns and verbs. We avoid generating new function words (e.g., new *wh*-word as a relative pronoun).

Nonce Sentence We have not dug into the difference between natural sentences and nonce sentences (Gulordava et al., 2018; Wei et al., 2021) that are grammatical but completely meaningless because we create additional training and evaluation instances with LLM, which tends to generate naturally plausible sentences. Nonce sentences are less plausible in human language acquisition but exclude semantic selectional-preferences cues (Gulordava et al., 2018; Goldberg, 2019). According to Section 7.1, there can be a trade-off between training language models in experimental settings that closely resemble natural human language acquisition and those that are strictly controlled. Future work can determine whether nonce sentences with indirect evidence differently affect linguistic generalization in language models.

Limited Model Size and Pretraining Data We use a small-scale language model and pretraining data in this work because we aim to find the differences from human inductive biases as much as possible. It is uncertain that the same trends as our work will appear in models of any size. Whether scaling laws apply to indirect data in accelerating model generalization would be an interesting future work.

Ethics Statement

There might be a possibility that the texts we used (Wikipedia) and the sentences generated by large language models are socially biased, despite their popular use in the NLP community.

Acknowledgments

We would like to express our gratitude to the anonymous reviewers who provided many insightful comments that have improved our paper. This work was supported by JSPS KAKENHI Grant Numbers JP21H05054, 22K17954, and 24KJ1700, and JST PRESTO Grant Numbers JPMJPR21C2 and JPMJPR20C4.

References

Ben Ambridge, Evan Kidd, Caroline F. Rowland, and Anna L. Theakston. 2015. The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, 42(2):239–273.

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

Noam Chomsky. 1993. *Lectures on Government and Binding*. De Gruyter Mouton, Berlin, New York.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Cara Su-Yi Leong and Tal Linzen. 2024. Testing learning hypotheses using neural networks by manipulating learning data.

Bingzhi Li, Guillaume Wisniewski, and Benoît Crabbé. 2023. Assessing the capacity of transformer to abstract syntactic representations: A contrastive analysis based on long-distance agreement. *Transactions of the Association for Computational Linguistics*, 11:18–33.

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.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

R. Thomas McCoy, Robert Frank, and Tal Linzen. 2020. Does syntax need to grow on trees? sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, 8:125–140.

Meta. 2024. The llama 3 herd of models.

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing aanns.

Aaron Mueller, Robert Frank, Tal Linzen, Luheng Wang, and Sebastian Schuster. 2022. Coloring the blank slate: Pre-training imparts a hierarchical inductive bias to sequence-to-sequence models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1352–1368, Dublin, Ireland. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4 technical report.

Abhinav Patil, Jaap Jumelet, Yu Ying Chiu, Andy Lapastora, Peter Shen, Lexie Wang, Clevis Willrich, and Shane Steinert-Threlkeld. 2024. Filtered corpus training (fict) shows that language models can generalize from indirect evidence.

Lisa S. Pearl and Benjamin Mis. 2016. The role of indirect positive evidence in syntactic acquisition: A look at anaphoric "one". *Language*, 92(1):1–30.

Ryokan Ri and Yoshimasa Tsuruoka. 2022. Pretraining with artificial language: Studying transferable knowledge in language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7302–7315, Dublin, Ireland. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Naoya Ueda, Masato Mita, Teruaki Oka, and Mamoru Komachi. 2024. Token-length bias in minimal-pair paradigm datasets. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16224–16236, Torino, Italia. ELRA and ICCL.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. 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. 2020a. BLiMP: A benchmark of linguistic minimal pairs for English. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 409–410, New York, New York. Association for Computational Linguistics.

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

Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. 2021. Frequency effects on syntactic rule learning in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 932–948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jennifer C. White and Ryan Cotterell. 2021. Examining the inductive bias of neural language models with artificial languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 454–463, Online. Association for Computational Linguistics.

Charles Yu, Ryan Sie, Nicolas Tedeschi, and Leon Bergen. 2020. Word frequency does not predict grammatical knowledge in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4040–4054, Online. Association for Computational Linguistics.

A Data generation

A.1 Linguistic phenomena

We employ seven linguistic phenomena, following (Warstadt et al., 2020b), to create training/evaluation instances. The linguistic phenomenon "transitive" is from "causative", "intransitive" is from "drop_arguement", "determiner-noun agreement" is from "determiner_noun_agreement_2",

```
Create 400 minimal sentence pairs, containing a grammatical and an ungrammatical sentence, following the template pair and rules.
Template pair:
[WUG] <singular transitive verb> herself.
[WUG] <singular transitive verb> himself.
Rules:
- You must include the lemma of <singular transitive verb> with a different initial letter and different final letter from the previous ones.
- Always use the female proper noun [WUG] with bracket[] and uppercase.
- You must include various auxiliary verbs and tenses in <singular transitive verb> with a different initial letter and different final letter from the previous ones.
- You often include negations in <singular transitive verb> if previous pairs did not contain ones.
- Do not include adverbs.
- Generate 400 pairs including numbering that starts from 1 and ends at 400.
Example:
[WUG] will hurt herself.
*[WUG] will hurt himself.
```

Figure 4: An example of prompt used to create evaluation examples.

"subject-verb agreement (V)" is from "regular_plural_subject_verb_agreement_1", and "subject-verb agreement (S)" is from "regular_plural_subject_verb_agreement_2".

A.2 Pretraining Data

We aim to pretrain the language models for 18 epochs while controlling the number of occurrences of target instances. To achieve this, we concatenate the pretraining data 18 times consecutively and randomly select where to inject each additional training instance.

A.3 Creating Data with LLM

The GPT-4 sometimes inconsistently generates sentences with hallucination; it generates the same sentence repeatedly and sometimes stops generating midway. To generate as many lexically diverse instances as possible, we prompt GPT-4 to avoid using the same lemma as in the previous instance. To get appropriate instances, we prompt the GPT-4 to generate double the number of instances[10], and then select the designated number of instances, avoiding duplicates. We adjust the percentage of sentences with negation words to 10–50%. The balanced instances contained 100 feminine and 100 masculine instances in ANA.GEN.AGR, 34 feminine singular and 33 masculine singular, 34 singular and 100 plural instances in ANA.NUM.AGR, 200 instances each in TRANSITIVE and INTRANSITIVE, 50 this, 50 that, 50 these and 50 those in D-N AGR. 100 singular and 100 plural each in S-V AGR.

A.4 Prompts

An example of prompts used to generate minimal sentence pair in anaphor gender agreement where a <wug#n> in the correct sentence is "herself" is shown in Figure 4. Another example is found in https://github.com/nii-cl/widet. We use `gpt-4-turbo` with top_p set to 1.0 and temperature set to 0.

[10]The number of instances generated based on the prompt can vary. Sometimes the output meets the specified quantity, while other times it may be fewer, potentially even less than half of the requested amount. If not enough instances are generated, we input instances from three steps earlier and generate additional instances to meet the requirements.

B Considering BLiMP Score Calculation

To select one sentence in each pair while evaluating, we calculate its sentence-level likelihood, referring to Warstadt et al. (2020a); Huebner et al. (2021). Conversely, Hu et al. (2020) argue that token-level likelihood comparisons, comparing the aggregate likelihood over a word like "herself" vs. a word like "himself", is a more precise evaluation than sentence-level probability. We consider the difference using the two phenomena as a case study.

Settings We compare the sentence-level likelihood used in this work with two types of score calculation; wug-level likelihood and reflexive-level likelihood. Given the sentence "<wug#n> has devoted herself/*himself," the reflexive-level likelihood compares the probabilities assigned to the reflexives "herself" and "himself." This is similar to the method used by Hu et al. (2020). The wug-level likelihood, on the other hand, compares the probabilities assigned to each pair of <wug#n>. Since we are using MLMs in our research, it is possible to adapt this for our calculations.

Results The score of language models calculated by the different score calculation methods are shown in Figure 5. Two phenomena are different trends. For anaphor gender agreement, the sentence-level and wug-level calculation methods show similar trends where the score increased gradually between 25 and 75 occurrences. The reflexive-level method does not show such a result but hits a plateau after 75 observations. For anaphor number agreement, the sentence-level and reflexive-level methods show similarities but the latter shows a bit more efficient learning than the former. The wug-level method does not show improvement until 100 observations. The results suggest that, in our limited setting, there are distinct

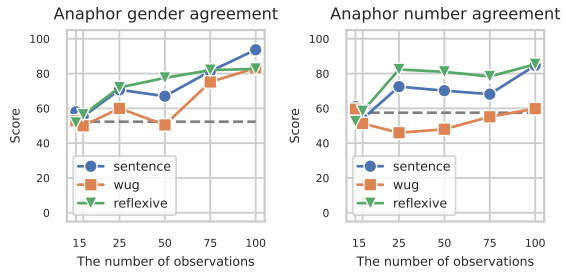

Figure 5: Model's score for three different score calculation methods

Model	architecture	roberta-base
	vocab size	9,600
	hidden size	512
	heads	8
	layers	8
	dropout	0.1
	layer norm eps	1e-12
	initializer range	0.02
Optimizer	algorithm	AdamW
	learning rates	2e-4
	betas	(0.9, 0.999)
	weight decay	0.0
Scheduler	type	linear
	warmup updates	24,000
Training	gradient accum.	4
	epoch	18
	batch size	16
	line by line	true
	NGPU	1

Table 6: Hyperparameters of the language models.

trends among the three methods. The sentence-level and reflexive-level methods each have their advantages depending on the language phenomena. More analyses of their difference are interesting for future work.

C Hyperparameters

Hyperparameters in our work are listed in Table 6.