CDL 2026

The 1st Workshop on Computational Developmental Linguistics (CDL)

Proceedings of the Workshop

July 3, 2026

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-428-6

Introduction

We are pleased to welcome you to the First Workshop on Computational Developmental Linguistics (CDL), held in conjunction with ACL 2026 in San Diego.
Computational developmental linguistics brings together questions about language acquisition, learning dynamics, adaptation, and change over time across human and artificial learners. The workshop is intended as a venue for researchers in machine learning, computational linguistics, cognitive science, and developmental psychology who are interested in studying how linguistic knowledge emerges, evolves, and adapts.

Background The first workshop on computational developmental linguistics (CDL) invites interdisciplinary contributions broadly in the topic of computational developmental linguistics. By computational developmental linguistics, we broadly refer to the studies of computational modeling of language acquisition and change over time, encompassing both individual learners, including humans or machines, throughout their lifetime and across populations. Particularly, we welcome submissions that study the developmental trajectories through which learners master phonology, lexicon, and syntax, as well as the idiolectal shifts that arise as speakers adapt to new communities, domains, or interlocutors. With the advent of neural language models (LMs), we now have large-scale systems that can be examined for learning dynamics, representational change, knowledge tracing, and language adaptation. Beyond static text corpora, interactive and multimodal training regimes make it possible to reverse-engineer developmental conditions, explore resource-efficient learning, and measure parallels and divergences between human and machine language acquisition. This rapidly growing area not only advances our understanding of how linguistic capabilities emerge and evolve in artificial systems, but also provides new potential for generating hypotheses about human language learning.

Scope and Goal This workshop aims to bridge the conversation between modern machine learning and developmental linguistics. We hope to draw inspiration from both fields, identifying similarities and differences in the (im)plausibilities of comparing LMs and human language learning, to motivate better LM engineering and more rigorous computational modeling of human language acquisition. Possible comparisons include (1) pretraining dynamics vs. first language acquisition: relating the data and scaling behavior of large models to the developmental timelines, stages, and milestones documented in human language learning; (2) continual learning and post-training vs. idiolect change and second language acquisition: exploring how ongoing adaptation, fine-tuning, or domain shifts in large models mirror semantic drift and individual language change in human speakers. By bringing together researchers from machine learning, computational linguistics, cognitive science, and developmental psychology, we aim to foster cross-disciplinary dialogue and establish frameworks for building computational models and performing computational analyses of language development, enabling scientifically rigorous comparisons between humans and machines without overly anthropomorphizing LMs.

Topics of Interest The scope and topics include, but are not limited to:

- Computational models for developmental linguistics. Models and formalisms for simulating first and second language acquisition, including data constraints, model architectures, training objectives, and frameworks for modeling idiolect change and semantic drift.
- Learning dynamics in pretraining and post-training of LMs. Behavioral and mechanistic interpretations of how linguistic and cognitive competence emerge during pretraining, and how it evolves through post-training (e.g., finetuning, alignment, and domain adaptation). In addition to text-only LMs, we encourage approaches that integrate multiple modalities and interactive learning, as these more closely mirror human language acquisition processes.

- Comparisons of developmental linguistics in humans and machines. Comparing the processes, constraints, and outcomes of language acquisition in humans and artificial learners; examining the possibilities, limitations, and methodological challenges of such comparisons; relating model learning trajectories to developmental timelines and milestones in humans; and conducting these analyses in a scientifically rigorous manner without over-anthropomorphizing LMs.
- Knowledge tracing and developmental diagnostics. Developing tools and methods to track the acquisition, transformation, and retention of linguistic knowledge in LMs, including phase transition detection, benchmarks, and probes of representational change.
- Applications in language education and clinical NLP. Applying insights from computational developmental linguistics to improve LM engineering, language tutoring systems, child-directed AI interaction, adaptive educational technology, and computational models of neurological communication disorders.

For this first edition of the workshop, we were particularly excited to see contributions spanning language modeling, speech and multimodal learning, child-directed data, interaction, semantic change, and the analysis of developmental trajectories in both human and machine language systems. The workshop also features invited talks and a panel discussion intended to support broad interdisciplinary exchange.
We thank the authors for their submissions and the reviewers for their time, care, and expertise. We are also grateful to the invited speakers, panelists, and organizers whose work made this workshop possible. We hope that CDL helps establish a durable meeting place for this emerging research area and supports future work at the intersection of developmental linguistics and modern machine learning.

Martin Ziqiao Ma, Emmy Liu, Jing Liu, Tyler A. Chang, Abdellah Fourtassi, Alex Warstadt, Michael Hahn, Weiwei Sun, and Freda Shi

Organizing Committee

Organizing Committee

Martin Ziqiao Ma, University of Michigan
Emmy Liu, Carnegie Mellon University
Jing Liu, École Normale Supérieure
Tyler A. Chang, Google DeepMind
Abdellah Fourtassi, Aix-Marseille University
Alex Warstadt, UC San Diego
Michael Hahn, Saarland University
Weiwei Sun, University of Cambridge
Freda Shi, University of Waterloo / Vector Institute

Program Committee

Area Chairs

Tyler A. Chang, Google DeepMind
Abdellah Fourtassi, Aix Marseille University
Michael Hahn, Saarland University
Emmy Liu, School of Computer Science, Carnegie Mellon University
Ziqiao Ma, Thinking Machines Lab
Freda Shi, Vector Institute and University of Waterloo

Reviewers

Bastian Bunzeck

Sarah C. Creel

Abdellah Fourtassi

Jugal Gajjar, Sadaf Ghaffari, Prachi Goyal

Nadia Ghezaiel Hammouda

Parin Rajesh Jhaveri

Shunsuke Kando, Garry Kuwanto

Siddharth Lall, Dan Le, Xue Li

Fred Mailhot, Akshata Kishore Moharir

Ruoxi Ning, Sergiu Nisioi

Udita Patel, Van-Thuy Phi

Mengyang Qiu

Suchir Salhan, Shivani Shekhar, Gyu-Ho Shin, Siyuan Song, Xiaonan Song, Dewang Sultania

Rohith Uppala

Tai Vu

Yixuan Wang, Ahmed Wez, Daniel Wurgaft

Aditya Yadavalli

Chen Zhang, Jiayi Zhang

Table of Contents

Linguistics Theory Meets LLM: Code-Switched Text Generation via Equivalence Constrained Large Language Models

Garry Kuwanto[1], **Chaitanya Agarwal**[2], **Genta Indra Winata**[†3*], **Derry Tanti Wijaya**[†1,4]
[1]Boston University [2]Deccan AI [3]Capital One [4]Monash University Indonesia
gkuwanto@bu.edu, chaitanya@deccan.ai
genta.winata@capitalone.com, derry.wijaya@monash.edu

Abstract

Code-switching is a common practice for millions of multilingual speakers but remains challenging for Large Language Models (LLMs). This paper investigates LLM capabilities in generating code-switched text, conducting extensive experiments across five diverse language pairs: English paired with Hindi, Tamil, Malayalam, and Indonesian, as well as Indonesian-Javanese. Our analysis, grounded in comprehensive human evaluations by native speakers, uncovers a directional asymmetry: LLMs consistently produce higher-quality (more accurate and fluent) code-switched text when prompted with a lower-resource language (e.g., Hindi, Tamil, Javanese) as the source, compared to when a higher-resource language (English, Indonesian) serves as the source. This asymmetry mirrors sociolinguistic patterns, particularly the Matrix Language Frame model, suggesting LLMs implicitly learn common code-switching structures from their training data where regional languages often form the grammatical base. Furthermore, we find that explicit linguistic guidance, applied through Equivalence Constraint Theory (ECT) to identify switching points, primarily benefits generation quality only in the less common, higher-resource-source direction where LLMs intrinsically struggle. These findings highlight a crucial interplay between the implicit linguistic knowledge captured by LLMs and the targeted utility of explicit linguistic constraints. We also introduce CSPREF, a pairwise preference dataset derived from our human evaluations, to facilitate future research in code-switching generation and evaluation.

1 Introduction

Bilingual and multilingual speakers frequently engage in code-switching, the phenomenon where speakers alternate between languages within a single discourse. The widespread of this linguistic phenomenon follows complex syntactic, semantic, and sociolinguistic patterns rather than occurring randomly or as an indication of the lack of language proficiency. Despite being so widespread, code-switching still remains a challenge to model effectively, creating significant barriers to building truly inclusive language technologies.

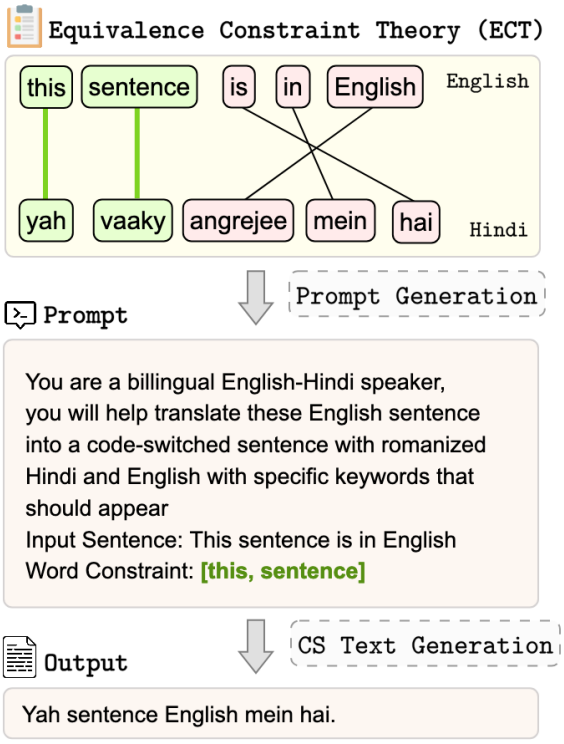


Figure 1: Example of CSPREF. The top panel shows the word-level alignment between English and Hindi. The middle panel displays the input to the LLM, including the original English sentence and word constraints derived from ECT. The bottom panel shows the resulting code-switched output sentence generated by the LLM.

The computational implementation of linguistic theories explaining code-switching constraints has proven difficult. Established frameworks like Equivalence Constraint Theory (Poplack, 1980) and Matrix Language Frame model (Myers-

* †The authors are senior authors.

Proceedings of the 1st Workshop on Computational Developmental Linguistics (CDL), pages 1–14

Scotton, 1997) offer theoretical explanations, but their application in NLP systems remains limited. Current approaches typically adopt either purely data-driven methods or focus on complex syntactic rules (Bhat et al., 2016), with few successfully bridging these approaches. Pratapa and Choudhury (2021) implements linguistic constraints using parse trees from parallel sentences. Similarly, Winata et al. (2019) applies linguistic constraints to generate synthetic code-switched text, finding that combining real and synthetic data improves performance for pretraining Language Models. Large Language Models (LLM) while demonstrating impressive cross-lingual capabilities, continue to struggle with generating natural code-switching (Winata et al., 2021; Zhang et al., 2023) often producing unnatural switching points due to insufficient training data of the same distribution.

To facilitate a comprehensive analysis of code-switching generation and evaluation, we conduct extensive experiments across five diverse language pairs (English-Hindi, English-Tamil, English-Malayalam, English-Indonesian, and Indonesian-Javanese). Our research reveals a striking directional asymmetry in code-switching quality that has significant implications for linguistic inclusion and technological equity. Our major contributions can be summarized as follows:

- We uncover a consistent pattern where LLMs generate higher-quality code-switched text when lower-resource languages serve as the source, without requiring explicit constraints. Conversely, when higher-resource languages serve as the source, additional linguistic guidance significantly improves quality.

- We demonstrate that this directional asymmetry mirrors sociolinguistic patterns observed in natural multilingual communication, where regional languages typically serve as the matrix language while global languages contribute embedded terms.

- We present a comprehensive evaluation dataset that prioritizes native speakers' judgments of code-switched text quality, offering a valuable resource for future research on evaluating code-switching across different language pairs and switching directions.

2 Linguistic Theories of Code-Switching

2.1 Code-Switching Patterns and Constraints

Code-switching, the alternation between two or more languages within a single discourse, represents a complex linguistic phenomenon observed across multilingual communities worldwide. Far from being random or indicative of language deficiency, code-switching follows systematic patterns governed by grammatical and sociocultural constraints (Poplack, 1980; Myers-Scotton, 1994; Muysken and ebrary, 2000.). Research has identified several common patterns, including inter-sentential switching (between sentences), intra-sentential switching (within a sentence), and tag-switching (insertion of a tag phrase) (Jan-Petter and Gumperz, 2007; Poplack, 1988).

The structural patterns of code-switching vary significantly across language pairs and communities. For instance, noun phrases are frequently switched in Spanish-English code-switching (Pfaff, 1979), while function words typically remain in the matrix language in Chinese-English mixing (Kamwangamalu and CHER-LENG, 1991). Several factors influence these patterns, including typological similarity between languages, lexical gaps, and discourse functions (Bullock and Toribio, 2009).

Sociolinguistic research has demonstrated that code-switching serves various communicative functions, including expressing solidarity, conveying nuanced meanings, establishing social identity, and filling lexical gaps (Gumperz, 1982; Myers-Scotton, 1993; Auer, 2013). Among bilingual communities with English as one of their languages, a common pattern emerges where the regional language often serves as the matrix language while English provides specific technical terminology, especially in domains like technology, education, and business (Bentahila, 1983; Wei, 2000).

2.2 Equivalence Constraint Theory

The Equivalence Constraint Theory (ECT), first proposed by Poplack (1980), represents one of the most influential frameworks for understanding the grammatical constraints on intra-sentential code-switching. The central premise of ECT is that code-switching is permissible only at points where the surface structures of the two languages align, meaning that switching does not violate the syntactic rules of either language involved.

More formally, ECT posits that code-switching

tends to occur at points where the word order requirements of both languages are satisfied simultaneously. If a potential switch point would create a structure that violates the grammar of either language, speakers typically avoid switching at that point. Sankoff (1998) formalized this constraint, stating that switching is prohibited if it generates a surface structure that would be ungrammatical in either language.

For example, in English-Spanish code-switching, switching between an adjective and noun is permissible when moving from English to Spanish but constrained in the opposite direction due to the differing adjective placement rules (English places adjectives before nouns, while Spanish typically places them after). Consider the sentence: "I bought a libro rojo" (I bought a red book), where switching occurs before "libro" (book). This switch respects both English syntax (determiner before noun) and Spanish syntax (noun before adjective). However, "I bought a rojo libro" would violate Spanish word order rules and is thus less likely to occur naturally (Poplack, 1978).

While ECT has proven effective in predicting many code-switching patterns, it faces limitations with typologically distant languages or in cases where one language strongly dominates as the matrix language (Muysken and ebrary, 2000.; Bhatt, 2013).

2.3 Matrix Language Frame Model

The Matrix Language Frame (MLF) model, developed by Myers-Scotton (1993, 1997), offers an alternative framework that addresses some limitations of ECT, particularly for language pairs with significant structural differences. This model distinguishes between the "matrix language" (ML), which provides the morphosyntactic framework, and the "embedded language" (EL), which contributes specific lexical items. According to the MLF model, the matrix language determines the overall grammatical structure, including word order and functional morphemes, while the embedded language contributes primarily content morphemes. Two key principles govern this interaction:

The Morpheme Order Principle: The order of morphemes must follow that of the matrix language. The System Morpheme Principle: System morphemes (functional elements like determiners, tense markers) must come from the matrix language.

This asymmetric relationship between languages helps explain why certain switching patterns occur more frequently than others. For instance, in many bilingual communities where a regional language interacts with English, the regional language typically serves as the matrix language, with English nouns and technical terminology embedded within the syntactic framework of the regional language (Bhatt, 2013; Sebba, 2009). The MLF model is particularly relevant for understanding directional asymmetry in code-switching patterns. It predicts that switching from the matrix language to the embedded language for content words (especially nouns and adjectives) is more common than the reverse direction (Myers-Scotton, 2002).

2.4 Directional Asymmetry in Natural Code-Switching

A significant yet often overlooked aspect of code-switching is its directional asymmetry across language pairs. Research has consistently demonstrated that the patterns, frequency, and constraints of code-switching differ depending on which language serves as the primary or matrix language (Gardner-Chloros, 2009; Sebba, 2009).

This asymmetry is particularly evident in post-colonial and globalized contexts where English interacts with regional languages. Studies across diverse communities—from Hindi-English in India (Kachru, 1978; Bhatt, 1997) to Swahili-English in East Africa (Myers-Scotton, 2002) and Spanish-English in the United States (Poplack, 1980)—reveal a consistent pattern: when the regional language serves as the matrix language, English lexical items are frequently embedded, especially for technical, academic, or professional domains. However, when English serves as the matrix language, embeddings from the regional language tend to be more limited and often serve cultural or identity-marking functions (Bullock and Toribio, 2009).

Several factors contribute to this asymmetry **Sociolinguistic status**: The relative prestige and domains of use for each language influence switching patterns (Myers-Scotton, 1993). **Lexical accessibility**: Terms may be more readily available or precise in one language than another (Heredia and Altarriba, 2001). **Processing constraints**: The cognitive mechanisms underlying language production may favor certain switching directions (Green and Abutalebi, 2013). **Communicative functions**: Different switching directions serve distinct discourse purposes (Gumperz, 1982).

The asymmetric nature of code-switching has important implications for computational modeling. Models trained primarily on data where one language consistently serves as the matrix may develop biases that affect their ability to generate natural code-switching in the opposite direction. This directional sensitivity suggests that linguistic constraints may be more critical for guiding generation in directions that are less represented in naturally occurring data.

Understanding these directional asymmetries is crucial for developing computational approaches that accurately reflect the complex patterns observed in natural code-switching across diverse multilingual communities.

3 Methodology

3.1 Experimental Design

To investigate directional asymmetry in code-switching, we designed a comprehensive evaluation framework examining code-switched text generation across five diverse language pairs: English-Hindi (en-hi), English-Tamil (en-ta), English-Malayalam (en-ml), English-Indonesian (en-id), and Indonesian-Javanese (id-jv). These pairs were selected to represent varying typological distances, resource availability, and sociolinguistic contexts. For each pair, we examined both possible directional flows: higher-resource language to lower-resource language (e.g., English→ Hindi) and lower-resource language to higher-resource language (e.g., Hindi→ English). Our experimental design focused on evaluating the quality of generated code-switched text under different conditions:

1. Direct Generation: Direct prompting of LLMs to produce code-switched text without explicit linguistic constraints

2. Linguistically-Guided Generation: Generation guided by constraints derived from linguistic theories, specifically Equivalence Constraint Theory (ECT)

3.2 Data Preparation

3.2.1 Obtaining Translations

We utilized existing parallel datasets to obtain high-quality translations across our target language pairs. For Hindi, we used the HinGE dataset (Srivastava and Singh, 2021), while Tamil and Malayalam translations came from the Samanantar dataset (Ramesh et al., 2022). For Indonesian-English and Indonesian-Javanese, we used the NusaX dataset (Winata et al., 2023b). These human-translated parallel sentences provided a reliable basis for our experiments.

To ensure consistency across experiments, we also generated translations using Llama3 8B, providing a controlled alternative to the human translations. This dual approach allowed us to assess whether any observed directional asymmetry was consistent across both human and machine translations.

3.2.2 Bitext Alignment

For each parallel sentence pair, we applied the GIZA++ tool (Och and Ney, 2003) to obtain word-level alignments between the source and target languages. These alignments were crucial for identifying potential code-switching points according to linguistic constraints.

3.3 Code-Switched Sentence Generation

3.3.1 Direct Generation

For Direct Generation, we prompted large language models to generate code-switched text without providing explicit linguistic constraints. The prompt simply instructed the model to create a code-switched version of the input sentence, incorporating elements from both languages naturally.

3.3.2 Linguistically-Constraint Generation

For the linguistically-guided approach, we incorporated switching constraints derived from Equivalence Constraint Theory. ECT posits that code-switching is natural at points where the word order rules of both languages align, avoiding violations of either language's syntax. We operationalized ECT by using word alignments to identify non-crossing alignment points as valid switching locations. For each input sentence, we constructed prompts that guided the model to generate the sentence with words that are acquired from

3.4 Evaluation Framework

To ensure a robust assessment of code-switching quality, we developed a comprehensive evaluation framework

3.4.1 Automatic Evaluation

Our primary evaluation methodology employs GPT-4o-mini as an automated assessor of code-switching quality. This approach demonstrates substantially higher correlation with human judgments

compared to traditional metrics, with Kendall's tau coefficients of 0.558 for accuracy and 0.514 for fluency (as detailed in 4.

We provide GPT-4o-mini with identical instructions as our human evaluators, asking it to rate generated sentences on discrete scales from 1 (lowest) to 3 (highest) for both accuracy and fluency. The prompt includes:

The original sentence in the first language (L1) The corresponding sentence in the second language (L2) The generated code-switched output

This structured approach allows GPT-4o-mini to perform consistent evaluations across different language pairs and generation methods, focusing specifically on meaning preservation (accuracy) and natural-sounding integration of languages (fluency).

3.4.2 Human Evaluation

We perform human evaluations with native bilingual speakers who actively code-switch in daily life. Indic language evaluators are recruited through DeccanAI (previously SoulAI), which ethically sources and trains annotators in India after assessing their English and native language proficiency. Indonesian evaluators are separately recruited. All annotators score the code-switched sentences for accuracy and fluency using established annotation guidelines.

This on 150 sample of inputs from dataset described in Section 4.3 with 18 different generation settings. Totaling 2700 sentence to rate for each language. In total, we conduct 24,300 human evaluations in total. For each code-switched sentence, we ask 3 unique evaluators to score the accuracy and fluency of the sentence on a discrete scale from 1 (lowest) to 3 (highest). While evaluating, the evaluators can see the parallel sentence pair (both languages) and the LLM generated code-switched sentence.

4 Experimental Setup

4.1 Models

We employ three distinct open-weight LLMs to assess the consistency of our findings across different architectures and training regimes:

- Aya 23 (8B): (Aryabumi et al., 2024) An LLM explicitly designed for multilingual tasks, trained on a diverse language corpus, making it potentially well-suited for code-switching.
- Llama 3 (8B): (Dubey et al., 2024) A widely-used model offering a balance between performance and computational efficiency within the Llama 3 series.
- Llama 3.1 (8B): An improved iteration of Llama 3 8B, incorporating refined training techniques and updated data, potentially offering enhanced capabilities in complex linguistic tasks like code-switching.

All experiments were conducted on a single NVIDIA L40 GPU equipped with 48GB of memory, ensuring consistency in the computational environment.

4.2 Language Pairs and Directions

Our investigation spans five language pairs, chosen to represent different language families, typological characteristics, and sociolinguistic contexts involving English and Indonesian as higher-resource languages: English-Hindi (en-hi), English-Tamil (en-ta), English-Malayalam (en-ml), English-Indonesian (en-id), Indonesian-Javanese (id-jv).

For each pair, we examine code-switched generation in two distinct directions to probe for asymmetry:

Higher-Resource to Code-Switched: Generation initiated from the higher-resource language (English for en-hi, en-ta, en-ml, en-id; Indonesian for id-jv), incorporating the lower-resource language.

Lower-Resource to Code-Switched: Generation initiated from the lower-resource language (Hindi, Tamil, Malayalam, Indonesian, or Javanese), incorporating the higher-resource language.

4.3 Datasets

Our experiments utilize five parallel corpora, each corresponding to a distinct language pair, as detailed in Table 1. For the human translations, we rely on the parallel sentences available in these datasets. In contrast, for the LLM translations we generate translations using the Llama3 8B model.

5 Results

5.1 Automatic Metrics

We first present the results using GPT-4o-mini as an automated evaluator for Accuracy (preserving meaning, GPT4o_a) and Fluency (naturalness of

Language Pair	Source	Size
Hindi-English (hi-en)	HinGE	2,766
Tamil-English (ta-en)	Samanantar (WAT 2020)	2,000
Malayalam-English (ml-en)	Samanantar (WAT 2020)	2,000
Indonesian-English (id-en)	NusaX	1,000
Indonesian-Javanese (id-en)	NusaX	1,000

Table 1: Summary of datasets used in the experiments, including source and size.

Method	Low	High	Δ
GPT4o$_a$			
Direct Generation	1.55	1.39	+0.26
Guided Gen. (Human Trans.)	1.58	1.47	+0.11
Guided Gen. (LLM Trans.)	1.59	1.47	+0.12
GPT4o$_f$			
Direct Generation	1.63	1.75	+0.12
Guided Gen. (Human Trans.)	1.53	1.67	+0.14
Guided Gen. (LLM Trans.)	1.53	1.64	+0.11

Table 2: Average GPT-4 evaluation of Accuracy (GPT4o$_a$) and Fluency (GPT4o$_f$) for the two direction generation across different methods. Scores are in the range of 1–3. Positive Δ shows that Low Resource is better than High Resource.

code-switching, GPT4o$_f$). While automatic metrics for code-switching are challenging, recent work suggests large models like GPT-4o can serve as reasonable proxies for certain quality aspects, especially when compared to traditional metrics (see Section 5.3). Table 2 shows the average scores across all five language pairs for each method and model, separated by the direction of generation.

The automatic evaluation scores in Table 2 provide initial evidence supporting directional asymmetry, with Direct Generation achieving higher accuracy (1.55 vs. 1.39) and fluency (1.75 vs. 1.63) when sourced from the lower-resource language. Interestingly, while the linguistically guided methods show some accuracy gains over direct generation, they appear to result in lower fluency scores according to GPT-4o across both directions. These preliminary automated results underscore the need for human evaluation to fully assess the nuances of code-switching quality, which we address next.

5.2 Human Evaluation Results

Given the limitations of automatic metrics for nuanced linguistic phenomena like code-switching fluency, we now turn to the human evaluation results presented in Table 3 as our primary basis for analysis. These scores, provided by native bilingual speakers actively using code-switching, offer crucial insights into the perceived accuracy and naturalness of the generated text.

lang	GPT4o$_a$	GPT4o$_f$	Human$_a$	Human$_f$
en-hi	**0.33**	**0.41**	**0.59**	0.53
en-ml	0.12	0.10	0.56	**0.63**
en-ta	0.07	0.12	0.45	0.39
id-jv	0.13	0.15	0.14	0.12
en-id	0.11	0.11	0.23	0.29

Table 3: Difference of Average between Low Resource and High Resource for Direct Generation. Difference Value is calculated by $\Delta_m = low_m - high_m$. All values are positive because Low Resource inputs consistently outperforms High Resource inputs. The higher the difference the higher the more prominent the asymmetry

Direct Generation The human evaluation results for direct generation reveal clear directional asymmetry across language pairs, with lower-resource languages consistently outperforming higher-resource languages as source inputs. English-Hindi shows the most pronounced asymmetry with human accuracy and fluency differences of 0.59 and 0.53 respectively, followed by English-Malayalam (0.56, 0.63) and English-Tamil (0.45, 0.39). The Indonesian-Javanese pair exhibits minimal asymmetry (0.14, 0.12), likely due to both languages occupying similar resource levels within their shared ecosystem. English-Indonesian shows moderate differences (0.23, 0.29). These patterns align with sociolinguistic observations that speakers naturally use their regional language as the grammatical foundation while incorporating English terms, rather than embedding regional elements within English grammatical structures.

5.3 Correlation Between Automatic and Human Metrics

To assess the reliability of automatic metrics for this task, we calculated the Kendall's Tau correlation between various automatic scores and the average human ratings (Accuracy and Fluency) on the human-evaluated subset (2,700 samples per language, aggregated). Table 4 shows these correlations

The results confirm findings from previous work (Guzmán et al., 2017; Winata et al., 2023a) that traditional metrics like BLEU and COMET show weak correlation with human judgments of code-switching quality, especially for fluency. No-

	Human$_a$	**Human$_f$**
Human$_a$	1.000	0.768
Human$_f$	0.768	1.000
GPT4o$_a$	0.558	0.504
GPT4o$_f$	0.540	0.514
COMET_avg	0.246	0.290
BLEU*	0.229	0.201

Table 4: Kendall's tau correlation scores between different automatic metrics and human evaluations for Human Accuracy and Human Fluency. *BLEU score can only be calculated for Hindi-English as there is no code-switched references for other language pairs.

Language Pair	**Total**	**Easy**	**Hard**
Hindi-English (hi-en)	17,460	9,621	7,839
Tamil-English (ta-en)	5,034	4,506	528
Malayalam-English (ml-en)	8,664	7,517	1,147
Indonesian-English (id-en)	22,430	2,394	20,036
Indonesian-Javanese (id-jv)	44,606	13,262	31,344

Table 5: Statistics of CSPREF for five language pairs (Hindi-English, Tamil-English, Malayalam-English, Indonesian-English, and Indonesian-Javanese). "Easy" pairs are defined as those with high rating differences, while "Hard" pairs are defined as those with low rating differences.

tably, our GPT-4o-mini based evaluations (GPT4o$_a$, GPT4o$_f$) demonstrate substantially higher correlation with human ratings (Tau around 0.51-0.56) compared to other automatic metrics. This supports its use as a more reliable proxy for large-scale automatic evaluation in this context, although human evaluation remains the gold standard.

5.4 Pairwise Preference Dataset

We construct CSPREF, a pairwise preference dataset using human ratings to evaluate the performance of different models in code-switched text generation. Each pair consists of two generated code-switched sentences compared based on their human-evaluated accuracy and fluency scores. To further analyze the performance, we split the dataset into **easy** and **hard** subsets. The **easy** subset includes pairs where the difference in human ratings is high (indicating a clear preference for one generated sentence), while the **hard** subset consists of pairs with minimal differences (indicating ambiguous preferences). Table 5 provides the statistics for the pairwise dataset across three languages: Hindi, Tamil, and Malayalam. We report the total number of pairs, as well as the breakdown into **easy** and **hard** subsets for each language pair.

6 Discussion

6.1 Explaining Directional Asymmetry

We found a directional asymmetry in code-switching: ECT-guided approaches significantly improve quality when higher-resource languages are the source, but offer minimal benefit with lower-resource source languages (like Indic or Austronesian languages). This aligns with the Matrix Language Frame model (Myers-Scotton, 1993), where one language provides the grammatical framework while another contributes lexical items. In multilingual communities, lower-resource regional languages typically serve as the matrix language with higher-resource languages providing embedded content words (Bhatt, 2013; Myers-Scotton, 2002). Our findings suggest LLMs have implicitly learned these natural patterns during pre-training, developing strong capabilities to generate code-switched text where lower-resource languages serve as the matrix. This matches sociolinguistic patterns in regions like India and Indonesia (Kachru, 1978; Nababan, 1991). Conversely, LLMs struggle with the less common pattern of higher-resource matrix languages with lower-resource embeddings unless explicitly guided by ECT constraints (Poplack, 1988).

6.2 Linguistic Constraints vs. Implicit Knowledge in LLMs

Our results reveal an interesting observation between explicit linguistic constraints and LLMs' implicit knowledge. In the lower-resource-to-higher-resource direction, LLMs generate fluent code-switched text without guidance, suggesting they've internalized common patterns from training data.

This contributes to the debate on whether LLMs truly learn linguistic rules or simply memorize patterns (McCoy et al., 2020; Linzen and Baroni, 2021). For frequent phenomena like code-switching from lower-resource to higher-resource languages, LLMs develop robust implicit knowledge aligning with linguistic theories like MLF. However, explicit constraints remain valuable for less common patterns.

GPT-4o-mini's effectiveness as an evaluator, with much higher correlation to human judgments than traditional metrics, reinforces this view. The model appears to have internalized not just genera-

tion capabilities but also human quality assessment criteria.

These findings point to a complementary relationship: linguistic theories like ECT provide valuable guidance for uncommon patterns, while LLMs excel at reproducing frequently observed phenomena without explicit constraints.

7 Related Work

Code-switching has been extensively studied from both linguistic and computational perspectives. Early linguistic theories, such as ECT (Poplack, 1980), establishes foundational principles for understanding syntactic boundaries in code-switching. Similarly, research by Joshi (1982) and Pfaff (1979) examine structural constraints and sentence processing in bilingual contexts. Recent computational approaches have adapted these theories into neural models. For instance, Winata et al. (2019) utilized ECT to generate synthetic data for training language models, while Gupta et al. (2020) employed pre-trained models to create code-switched text without explicit constraints. Pratapa and Choudhury (2021) utilized ECT to synthetically generate code-switched text by using the Dependency Tree. And Gupta et al. (2021) adopted a Machine Translation approach to the problem. Comprehensive survey by Sitaram et al. (2019); Winata et al. (2023a) outline the computational challenges and advancements in code-switching research.

Evaluation benchmarks, such as LinCE (Aguilar et al., 2020) and GLUECoS (Khanuja et al., 2020) have standardized model assessments across diverse tasks. Recent studies have also investigated automatic metrics for code-switching (Guzmán et al., 2017) and explored the use of LLMs in understanding code-switched text (De Leon et al., 2024), and also generating (Yong et al., 2023). In this context, our work builds on these foundations by integrating linguistic constraints into LLM-based generation, addressing existing limitations in fluency and accuracy evaluation.

8 Future Work

Future work could address these limitations through several avenues. Using the CSPREF dataset to fine-tune models specifically for code-switching generation could potentially improve performance without explicit constraints. Combining ECT with other theories like MLF might yield a more comprehensive approach to constraint-guided generation. Adapting generation to match individual code-switching patterns based on demographic and language proficiency factors represents another promising direction Figure 2. Creating specialized automatic metrics that better correlate with human judgments of code-switching quality would also significantly advance the field. The asymmetric results also suggest an intriguing direction for future work: exploring whether models can be trained to recognize the "matrix language" in a given context and automatically apply appropriate constraints based on the direction of generation.

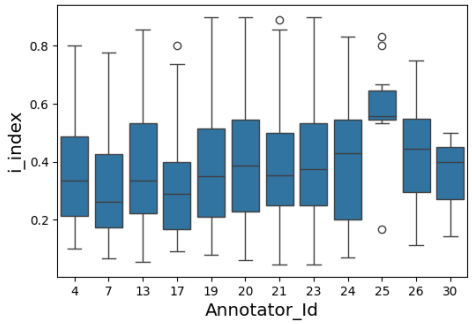


Figure 2: Distribution of I-index across annotators, representing the probability of code-switching at any given token. The I-index indicates the proportion of switch points relative to language-dependent tokens in the corpus. The variability in switching preferences among annotators highlights the individual differences in their judgment of fluency, suggesting that demographic factors may play a role in code-switching evaluation.

9 Conclusion

In this work,we investigated the capabilities of Large Language Models in generating code-switched text. Our evaluation across five language pairs revealed a striking directional asymmetry. Our results consistently show that models generate substantially more accurate and fluent code-switched text when prompted with a lower-resource language as the source, compared to when starting with a higher-resource language like English or Indonesian.

The asymmetry aligns with sociolinguistic patterns observed in natural code-switching and suggests that LLMs have implicitly learned common code-switching patterns during pre-training. Our findings demonstrate that linguistic theory and data-driven approaches can complement each other, with explicit constraints providing valuable guidance for less common code-switching patterns.

Our result underscore the importance of con-

sidering directionality and sociolinguistic context when developing and evaluating multilingual models. The challenges observed in automatic evaluation further emphasize the continued necessity of human judgment, a need we aimed to support by creating the CSPREF pairwise preference dataset. Ultimately, this work contributes to a more nuanced understanding of LLM capabilities and limitations in handling code-switching, paving the way for more linguistically informed and culturally aware language technologies.

Limitations

While our study provides valuable insights into code-switching generation, several limitations warrant discussion. Our language coverage, though spanning Indo-European, Dravidian, and Austronesian language families, could be expanded to include other language families, particularly tonal languages and those with substantially different writing systems, to strengthen our findings. The approach also focuses primarily on syntactic constraints and does not fully account for the sociolinguistic and pragmatic factors that influence code-switching in natural settings. Despite using GPT-4o-mini evaluation and human judgments, we still lack specialized metrics designed specifically for code-switching quality assessment. Additionally, our experiments are limited to relatively small open-source models; larger models might show different patterns or capabilities.

Ethics Statement

All aspects of this research were reviewed and approved by the Institutional Review Board of our organization. Data collection was conducted by DeccanAI for the Hindi, Tamil, and Malayalam evaluations. We compensate human evaluators INR 110 for every 18 sentences they evaluate, which typically takes around 20 minutes. This results in an effective pay rate of INR 330 per hour. The human evaluators work entirely remotely and interact with DeccanAI through their web platform. All evaluators are native speakers of the respective lower-resource languages they assess and are proficient in English. Their language proficiency is evaluated through custom online tests. Most evaluators come from major cities in India where these native languages are spoken and frequently engage in code-switched dialogues. DeccanAI provides training for the evaluators to ensure they are well-calibrated with the annotation guidelines.

For the Indonesian-Javanese language pair, annotators were recruited separately through our Indonesian university partners. These evaluators were compensated at a rate of IDR 2,000,000, in line with local research assistant compensation rates. All Indonesian annotators were native speakers of both Indonesian and Javanese, with most coming from various cities across Central and East Java, representing different dialectal backgrounds.

References

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. Lince: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1803–1813.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, and 1 others. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Peter Auer. 2013. *Code-switching in conversation: Language, interaction and identity*. Routledge.

Abdelali Bentahila. 1983. Motivations for code-switching among arabic-french bilinguals in morocco. *Language & communication*, 3(3):233–243.

Gayatri Bhat, Monojit Choudhury, and Kalika Bali. 2016. Grammatical constraints on intra-sentential code-switching: From theories to working models. *Preprint*, arXiv:1612.04538.

Rakesh M Bhatt. 2013. Optimization in bilingual language use. *Bilingualism: Language and cognition*, 16(4):740–742.

Rakesh Mohan Bhatt. 1997. Code-switching, constraints, and optimal grammars. *Lingua*, 102(4):223–251.

Barbara E Bullock and Almeida Jacqueline Ed Toribio. 2009. *The Cambridge handbook of linguistic code-switching*. Cambridge university press.

Frances Adriana Laureano De Leon, Harish Tayyar Madabushi, and Mark Lee. 2024. Code-mixed probes show how pre-trained models generalise on code-switched text. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3457–3468.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Penelope Gardner-Chloros. 2009. *Code-switching*. Cambridge university press.

David W Green and Jubin Abutalebi. 2013. Language control in bilinguals: The adaptive control hypothesis. *Journal of cognitive psychology*, 25(5):515–530.

John J Gumperz. 1982. *Discourse strategies*. 1. Cambridge University Press.

Abhirut Gupta, Aditya Vavre, and Sunita Sarawagi. 2021. Training data augmentation for code-mixed translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5760–5766.

Deepak Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A semi-supervised approach to generate the code-mixed text using pre-trained encoder and transfer learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2267–2280.

Gualberto A Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara E Bullock, and Almeida Jacqueline Toribio. 2017. Metrics for modeling code-switching across corpora. In *Interspeech*, pages 67–71.

Roberto R Heredia and Jeanette Altarriba. 2001. Bilingual language mixing: Why do bilinguals code-switch? *Current directions in psychological science*, 10(5):164–168.

Blom Jan-Petter and John J. Gumperz. 2007. Social meaning in linguistic structure: code-switching in norway. In *The Bilingualism Reader*.

Aravind Joshi. 1982. Processing of sentences with intra-sentential code-switching. In *Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics*.

Braj B Kachru. 1978. Toward structuring code-mixing: An indian perspective. *International Journal of the Sociology of Language*.

Nkonko M Kamwangamalu and LEE CHER-LENG. 1991. Chinese-english code-mixing: a case of matrix language assignment. *World Englishes*, 10(3):247–261.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. Gluecos: An evaluation benchmark for code-switched nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7(1):195–212.

R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2020. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 217–227, Online. Association for Computational Linguistics.

Pieter. Muysken and Inc. ebrary. 2000. *Bilingual speech*. Cambridge University Press,, Cambridge, UK ;.

Carol Myers-Scotton. 1993. *Social motivations for codeswitching: Evidence from Africa*. Oxford University Press.

Carol Myers-Scotton. 1994. Social motivations for codeswitching. evidence from africa. *Multilingua-Journal of Interlanguage Communication*, 13(4):387–424.

Carol Myers-Scotton. 1997. *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

Carol Myers-Scotton. 2002. *Contact linguistics: Bilingual encounters and grammatical outcomes*. Oxford University Press, USA.

PWJ Nababan. 1991. Language in education: The case of indonesia. *International review of education*, 37:115–131.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Carol W Pfaff. 1979. Constraints on language mixing: Intrasentential code-switching and borrowing in spanish/english. *Language*, pages 291–318.

Shana Poplack. 1978. *Syntactic structure and social function of code-switching*, volume 2. Centro de Estudios Puertorriqueños,[City University of New York].

Shana Poplack. 1980. Sometimes i'll start a sentence in spanish y termino en espanol: toward a typology of code-switching1. *Linguistics*, 18(7-8):581–618.

Shana Poplack. 1988. Contrasting patterns of code-switching in two communities. *Codeswitching: Anthropological and Sociolinguistic Perspectives. New York: Mouton de Gruyter*, pages 215–244.

Adithya Pratapa and Monojit Choudhury. 2021. Comparing grammatical theories of code-mixing. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 158–167.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan Ak, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, and 1 others. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

David Sankoff. 1998. A formal production-based explanation of the facts of code-switching. *Bilingualism: language and cognition*, 1(1):39–50.

Mark Sebba. 2009. Sociolinguistic approaches to writing systems research. *Writing systems research*, 1(1):35–49.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W Black. 2019. A survey of code-switched speech and language processing. *arXiv preprint arXiv:1904.00784*.

Vivek Srivastava and Mayank Singh. 2021. Hinge: A dataset for generation and evaluation of code-mixed hinglish text. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 200–208.

Lee Wei. 2000. *The bilingualism reader*, volume 11. Routledge London.

Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023a. The decades progress on code-switching research in nlp: A systematic survey on trends and challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023b. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in code-switching? In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 142–153.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2019. Code-switched language models using neural based synthetic data from parallel sentences. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 271–280.

Zheng Xin Yong, Ruochen Zhang, Jessica Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Lintang Sutawika, Jan Christian Blaise Cruz, and 1 others. 2023. Prompting multilingual large language models to generate code-mixed texts: The case of south east asian languages. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 43–63.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Aji. 2023. Multilingual large language models are not (yet) code-switchers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582.

A Relaxed Equivalence Constraint Theory (ECT)

Winata et al. (2019) apply ECT by simplifying sentences in terms of a linear grammatical structure and allowing lexical substitution on *non-crossing alignments* between parallel sentences (e.g., lexical substitution between "sentence" and "vaaky" in Figure 1). Denoting L_1 as the source language and L_2 as the target language, given a sentence in L_1 comprising an array of words $u_t = a_1, a_2, ..., a_m$ and a corresponding sentence in L_2 comprising an array of words $v_t = b_1, b_2, ..., b_m$, the alignment between a_i and b_i does not satisfy the constraint if there exists a pair a_j and b_j such that ($a_i < a_j$ and $b_i > b_j$) or ($a_i > a_j$ and $b_i < b_j$). If a switch occurs at this point, it alters the grammatical order in both languages, rendering the switch unacceptable. During the generation step, we permit *any* switches that do not violate this constraint.

This relaxation allows for greater flexibility in identifying potential switching points, accommodating the complexities of real-world code-switching patterns while maintaining grammatical coherence. Our implementation expands on the linear grammatical structure and non-crossing alignment criteria, introducing additional flexibility to capture a broader range of code-switching phenomena. In the following section, we outline our approach to implementing the relaxed ECT for identifying switching points, developing code-switched sentence generation techniques, and establishing a comprehensive evaluation framework.

B Switching Point Algorithm

The algorithm for the process of getting valid switching points is as described in Algorithm 1

C Prompt Details

In Table 6 we present the specific prompts used across different methods in our code-switching experiments. The Translate prompt is used when we generate the translations to get alignment. The Direct Generation prompt is used when we evaluate LLM to generate codeswitching. The ECT prompt is used for both Human translated and Machine

Algorithm 1: Identification of Valid Switching Points

```
Result: List of valid switching points
GetValidSwitchingPoints(pairs)
valid_pairs ← [ ];
for i ← 1 to length(pairs) do
    valid ← true;
    for j ← 1 to length(pairs) do
        (a_i, b_i) ← pairs[i];
        (a_j, b_j) ← pairs[j];
        if (a_i < a_j and b_i > b_j) or
         (a_i > a_j and b_i < b_j) then
            valid ← false;
            break;
        end
    end
    if valid then
        Append pairs[i] to valid_pairs;
    end
end
return valid_pairs;
```

Translated Linguistically-Guided Genration. The GPT Eval prompt defines the structure for evaluating code-switching output based on accuracy and fluency.

D Human Evaluation

D.1 Annotation Guidelines

The following guidelines are provided to human evaluators to assess the model's responses. Evaluators rate the generated sentences based on two criteria: **Accuracy** and **Fluency**. The original sentence is in English, Indian local languages (Hindi, Malayalam, and Tamil), and evaluators must adhere to the rubrics outlined below.

D.1.1 General Guidelines

- **MUST**: Be objective while rating the responses.
- **MUST**: Strictly follow the rubrics for Accuracy and Fluency evaluation.
- Score each criterion on a scale from 1 to 3, where 1 is the lowest and 3 is the highest.
- Ignore formatting, and any additional explanatory text generated by the language model. Only focus on meaning and context.
- If the model fails to generate a response, assign a score of 1 for both Accuracy and Fluency.

D.1.2 Accuracy

Accuracy measures how well the generated sentence preserves the meaning and information of the original sentence and whether the code-switched terms are used correctly. The scores are as follows:

- **1. Low Accuracy**:
  - Significant deviations from the original meaning.
  - Key information is missing, altered, or repeated redundantly.
  - Code-switched terms are incorrect or inappropriate.
  - Introduces new information not present in the original sentence.
  - Key details are altered or repeated redundantly.
- **2. Moderate Accuracy**:
  - Minor deviations from the original meaning.
  - Most key information is present but may have slight errors.
  - Most code-switched terms are appropriate but with minor mistakes.
- **3. High Accuracy**:
  - Preserves the original meaning fully.
  - All key information is present and correct.
  - Code-switched terms are accurate and appropriately used.

D.1.3 Fluency

Fluency measures how natural and easy to understand the generated sentence is, considering grammar, syntax, and the smooth integration of code-switching. The scores are as follows:

- **1. Low Fluency**:
  - The sentence is difficult to understand or awkward.
  - Poor grammar or syntax in either language.
  - Code-switching disrupts the flow of the sentence.

Method	Prompt
Translate	Translate the following `lang1` sentence to `lang2`: `<Input Sentence>`
Baseline	You are a Bilingual `lang1-lang2` speaker, you will help translate these `lang1` sentences into a code-mixed sentence with Romanized `lang2` and `lang1` `<Input Sentence>`
ECT Prompt	You are a Bilingual `lang1-lang2` speaker, you will help translate these `lang1` sentences into a code-mixed sentence with Romanized `lang2` and `lang1` with specific keywords that should to appear. `<Input Sentence>` Words wanted: `<List of Words>`
GPT Eval	You are provided with triplets of sentences. The first two sentence in each triplet is the original monolingual sentences. The second sentence is a generated code-switched sentence. Your task is to evaluate the generated sentence based on two criteria: Accuracy and Fluency. You will score each criterion on a scale from 1 to 3, where 1 is the lowest and 3 is the highest. When evaluating the generated sentences, focus on the content and meaning. Ignore any extra formatting, alignment artifacts, or additional explanatory text. Judge the sentence to determine its accuracy and fluency. original_l1: `<Original Lang1 Sentence>` original_l2: `<Original Lang2 Sentence>` generated: `<Code Switched Sentence>`

Table 6: Prompts used in our experiment.

- **2. Moderate Fluency**:
  - The sentence is understandable but may have awkward or unnatural phrasing.
  - Acceptable grammar and syntax in both languages.
  - Code-switching is somewhat smooth but not perfectly integrated.
- **3. High Fluency**:
  - The sentence is natural and easy to understand.
  - Good grammar and syntax in both languages.
  - Code-switching is smooth and seamless, enhancing the sentence flow.

D.2 Detailed Values

Table 7 shows the mean scores of GPT and Human judgements for all the combinations of experiments that we did.

D.3 Inter Annotator Agreement

As seen in Table 8, the inter-annotator agreement, measured by Krippendorff's alpha, reveals varying levels of consensus across languages, with the highest agreement for Hindi. While Fluency is generally lower, this is expected as Fluency is more of a subjective measure.

For Indonesian-Javanese, we observed lower agreement scores compared to other language pairs. This can be attributed to our annotators coming from different cities across Java, where regional variations in Javanese dialects led to different interpretations of certain common words. These dialectal differences affected how annotators judged the appropriateness of specific code-switched terms, particularly when evaluating fluency in contexts where regional expressions were used.

Throughout the evaluation process, we continuously monitored the quality of annotations by measuring inter-annotator agreement at regular intervals. If the agreement metric indicated significant divergence in scores, particularly when individual annotators' ratings deviated notably from the group consensus, we conducted alignment meetings. These meetings were used to clarify the guidelines and ensure a consistent understanding of the evaluation criteria among the annotators. During

		GPT4o$_a$		GPT4o$_f$		Human$_a$		Human$_f$	
	direction	high	low	high	low	high	low	high	low
lang	method								
en-hi	Direct Generation	1.69	2.02	1.75	2.14	1.75	2.33	1.79	2.32
	Guided Gen. (Human Trans.)	1.63	1.90	1.70	1.98	1.72	2.21	1.73	2.12
	Guided Gen. (LLM Trans.)	1.65	1.86	1.71	1.96	1.75	2.19	1.75	2.14
en-ml	Direct Generation	1.21	1.33	1.39	1.43	1.15	1.81	1.15	1.78
	Guided Gen. (Human Trans.)	1.36	1.27	1.46	1.38	1.17	1.72	1.16	1.66
	Guided Gen. (LLM Trans.)	1.38	1.33	1.48	1.40	1.16	1.66	1.15	1.61
en-ta	Direct Generation	1.26	1.33	1.31	1.43	1.11	1.56	1.15	1.54
	Guided Gen. (Human Trans.)	1.39	1.36	1.46	1.52	1.17	1.35	1.19	1.38
	Guided Gen. (LLM Trans.)	1.29	1.35	1.34	1.49	1.15	1.35	1.17	1.38
id-jv	Direct Generation	1.43	1.86	1.82	1.97	2.11	2.52	2.01	2.13
	Guided Gen. (Human Trans.)	1.48	1.56	1.62	1.69	2.27	2.40	1.97	1.99
	Guided Gen. (LLM Trans.)	1.48	1.52	1.66	1.62	2.29	2.36	1.99	1.97
en-id	Direct Generation	1.48	1.59	1.66	1.77	2.44	2.67	2.18	2.49
	Guided Gen. (Human Trans.)	1.51	1.56	1.67	1.77	2.53	2.56	2.20	2.26
	Guided Gen. (LLM Trans.)	1.58	1.61	1.77	1.74	2.47	2.52	2.17	2.21

Table 7: Mean scores of Human Accuracy (Human$_a$), Human Fluency (Human$_f$), and GPT4-based evaluations (GPT4o$_a$ for Accuracy and GPT4o$_f$ for Fluency). Scores are grouped by translation direction (Higher Resource to Lower Reousrce and vice versa.)

these sessions, any inconsistencies were discussed and resolved to improve consistency, especially in subjective aspects like Fluency. This iterative process helped ensure the reliability of the final evaluations and minimized discrepancies in the ratings.

Language	**Fluency**	**Accuracy**
Tamil-English	0.321	0.445
Malayalam-English	0.405	0.423
Hindi-English	0.646	0.720
Indonesian-English	0.535	0.606
Indonesian-Javanese	0.274	0.317

Table 8: Krippendorff's alpha scores for inter-annotator agreement on Fluency and Accuracy across Tamil, Malayalam, and Hindi.

Do Structural Priors Help Neural Language Models Learn Grammar? Evidence from Child-Scale Data

Jon-Paul Cacioli
Independent Researcher
Melbourne, Australia
synthium@hotmail.com

Abstract

We show that structural grammatical priors produce targeted, linguistically specific effects on grammatical learning: improving filler-gap dependencies — which require long-distance hierarchical tracking — by 9–13 percentage points beyond structural regularisation alone ($d = 2.41$–2.82), while damaging locally cued phenomena regardless of whether the grammar is real or random. This phenomenon-specificity, revealed by a random grammar control, suggests the right question is not whether structural priors help, but for which constructions and why. We test this by augmenting BabyBERTa (7.4M parameters) with a differentiable PCFG auxiliary loss derived from Minimalist Grammar, trained on AO-CHILDES (893K sentences of child-directed speech). In a pre-registered study of 190 experimental runs spanning 7 constraint strengths, 3 data scales, 5 random seeds, and 3 independent lexicon permutations, our confirmatory hypotheses about overall accuracy and sample efficiency are falsified. However, a random grammar control ($n = 15$ runs per condition; three independent lexicon permutations) reveals that linguistically accurate category assignments specifically drive filler-gap gains: real grammar outperforms both a structurally equivalent random grammar and the no-grammar baseline, while both conditions equally damage subject-verb agreement. These results show that structural priors function as targeted interventions rather than global boosters: they help specifically the constructions, specifically long-distance dependencies, whose computational demands align with what phrase-structure representations encode. We release code and pre-registered materials.[1]

[1]Pre-registration: https://osf.io/5rz9w/. Code: https://github.com/synthiumjp/neurosym-grammar. Hypotheses were registered after smoke-testing confirmed code correctness, but before the main experimental grid was run. All results reported regardless of outcome.

1 Introduction

The BabyLM Challenge (Warstadt et al., 2023, 2024) has demonstrated that small language models trained on developmentally plausible data can acquire surprising amounts of grammatical knowledge. BabyBERTa (Huebner et al., 2021) achieves grammatical competence comparable to RoBERTa-base using 6,000× fewer words and 15× fewer parameters, establishing a strong baseline for what distributional learning alone can accomplish at child scale. These results raise a question central to both computational linguistics and acquisition research: can explicit structural priors improve learning beyond what distributional statistics provide?

Cognitive scientists have long debated whether grammatical knowledge can be acquired from linguistic input alone or requires innate structural biases (Chomsky, 1965; Pullum and Scholz, 2002; Clark and Lappin, 2010; Pearl, 2022). Computational models offer a way to test these claims empirically: if adding structural priors to a neural learner measurably improves grammatical generalisation, this constitutes evidence that such priors are *useful*, regardless of whether they are innate (Warstadt and Bowman, 2022). Recent work has shown that neural language models can learn aspects of filler-gap dependencies and island constraints from data alone (Wilcox et al., 2024), but these models train on orders of magnitude more data than children receive. Whether structural priors help at genuinely developmental scales remains untested.

Despite advances in neurosymbolic grammar induction (Kim et al., 2019; Yang et al., 2021; Park and Kim, 2025), developmentally plausible language modelling (Huebner et al., 2021; Warstadt et al., 2023), and recent work integrating Minimalist Grammar into BabyLM training (Chesi et al., 2024), no prior work has combined explicit grammar constraints at child-scale data volumes with a random grammar control that decomposes struc-

Proceedings of the 1st Workshop on Computational Developmental Linguistics (CDL), pages 15–26

tural regularisation from linguistic content. We address this gap with four contributions:

1. A **neurosymbolic architecture** combining BabyBERTa with a differentiable PCFG auxiliary loss, where Minimalist Grammar-inspired rules shape learning through gradient-based structural supervision.

2. A **comprehensive evaluation** across 13 grammatical phenomena, 7 constraint strengths, 3 data scales (25%, 50%, 100% of AO-CHILDES), and 5 random seeds, totalling 130 experimental runs (plus 60 random grammar control runs).

3. A **random grammar control** with three independent lexicon permutations (60 additional runs) that decomposes observed effects into structural regularisation (from the CKY computation itself) versus linguistically specific content (from accurate category assignments).

4. **Falsified confirmatory hypotheses** reported transparently alongside informative exploratory findings from the random grammar analysis.

Our key finding is that the question "do structural priors help?" does not have a uniform answer across the grammar. Correct structural supervision specifically improves filler-gap dependencies beyond what structural regularisation alone provides, with large effect sizes ($d = 2.41$–2.82), while failing to improve — and in some cases damaging — locally cued phenomena. This phenomenon-specific pattern suggests that structural priors are most valuable for constructions requiring long-distance hierarchical tracking, the class of dependencies that motivates much of the debate around innate linguistic knowledge (Chomsky, 1965; Wilcox et al., 2024).

2 Method

2.1 Architecture

We extend BabyBERTa (Huebner et al., 2021), a small RoBERTa variant (7.4M parameters, 8 layers, 8 attention heads, 256 hidden dimensions) optimised for child-scale data. Training uses a combined loss:

$$\mathcal{L} = \mathcal{L}_{\text{MLM}} + \lambda \cdot \mathcal{L}_{\text{grammar}} \quad (1)$$

where $\mathcal{L}_{\text{MLM}}$ is the standard masked language modelling objective and $\mathcal{L}_{\text{grammar}}$ is the negative log-probability of the input under a probabilistic context-free grammar (PCFG), computed via a differentiable inside algorithm. The hyperparameter λ controls constraint strength.

The grammar loss operates through a **SoftLexicon** routing mechanism: a fixed, non-learnable matrix $M \in \mathbb{R}^{|V| \times |C|}$ maps from the model's token probability distribution to grammar category probabilities, where $M[\text{token}, \text{category}] = 1.0$ if the lexicon maps that token to that category. Concretely, given softmax output $p \in \mathbb{R}^{|V|}$, category probabilities are computed as $c = M^\top p$, yielding a distribution over grammatical categories that is then fed into the CKY inside algorithm. This matrix is determined entirely by the hand-crafted lexicon and contains no trainable parameters, ensuring that the grammar acts as a prior rather than an additional learned component. This approach contrasts with Chesi et al. (2024), who integrate Minimalist Grammar constraints into BabyLM training by modifying the gating mechanisms of RNNs rather than adding a differentiable auxiliary loss to a transformer. During training, the gradient chain flows from model logits through softmax token probabilities, through the category matrix, into the Cocke–Kasami–Younger (CKY) inside algorithm, and back, providing structural supervision without requiring discrete parsing decisions. We repurpose the CKY inside algorithm — traditionally used for parsing — as a differentiable scoring function: rather than finding the best parse, it marginalises over all possible parses to compute the total probability of the sentence under the grammar, and this scalar probability serves as the loss term.

Because the softmax produces non-zero probabilities for all tokens, gradient always flows; the lexicon determines the *direction* of this gradient by specifying which category each token is routed toward. For computational efficiency, grammar loss is computed every 4 training steps on 25% of each batch, with sentences truncated to 12 tokens (CKY complexity is $\mathcal{O}(n^3)$). This truncation is developmentally plausible: children's early utterances are predominantly short (Brown, 1973; Huebner et al., 2021), and the AO-CHILDES corpus has a median sentence length of 7 tokens.

2.2 Grammar

The grammar is a PCFG inspired by Minimalist Grammar (Stabler, 1997), implemented in Chom-

sky Normal Form. It comprises 17 nonterminal categories, 12 terminal categories, and 16 production rules encoding English phrase structure (e.g., TP $\rightarrow$ NP VP; VP $\rightarrow$ V NP; CP $\rightarrow$ C TP). The lexicon maps 7,230 word types to grammatical categories with associated features. The grammar directly encodes six of the thirteen evaluated phenomena: subject-verb agreement, determiner-noun agreement, argument structure, filler-gap dependencies, island effects, and local attractor configurations. The remaining seven phenomena (binding, NPI licensing, quantifiers, anaphor agreement, ellipsis, irregular forms, case) are not encoded in the grammar, providing a natural test of constraint specificity.

While inspired by Minimalist Grammar, our PCFG is a phrase-structure approximation rather than a full MG implementation: it encodes hierarchical constituent structure and basic filler-gap configurations through the CP $\rightarrow$ C TP rule, but cannot represent feature-checking, Agree relations, or the negative island constraints that require specifying where movement is prohibited. This scope is important for interpreting our results; benefits are expected only for phenomena whose structural properties fall within what the grammar can express.

2.3 Training Data

We use AO-CHILDES (Huebner et al., 2021), aggregating child-directed speech from the CHILDES database (MacWhinney, 2000) for children aged 1–6. The corpus contains 893,989 sentences ($\sim$5M words). To test sample efficiency, we create pregenerated random subsamples at 25% (223K sentences) and 50% (447K sentences). The grammar auxiliary loss is computed on sentences truncated to 12 sub-word tokens for computational tractability (the CKY inside algorithm is $\mathcal{O}(n^3)$). Given AO-CHILDES's mean sentence length of 7.3 subtokens (Huebner et al., 2021), this limit affects only $\sim$15–17% of training sentences, preserving the short conversational utterances characteristic of child-directed speech. The grammar constraint nonetheless generalises to full-length Zorro test items at evaluation time, indicating that the structural prior shapes learned representations rather than requiring explicit parsing of all training sentences.

2.4 Random Grammar Control

To distinguish linguistic content from structural regularisation, we construct a random grammar control by permuting the lexicon's word-to-category mappings. This preserves the exact category frequency distribution, the full CKY computation with identical algorithmic structure, and the same number of trainable gradient steps. It destroys only the linguistic accuracy of category assignments. Any performance difference between real and random grammar therefore isolates the contribution of linguistically accurate structural supervision, controlling for the regularisation effect of the CKY computation itself.

We use three independent permutations (seeds 99, 2026, and 31415), yielding $n = 15$ random-grammar runs per condition and tighter confidence intervals than a single permutation would provide. The three permutations produce highly consistent results (inter-permutation SD $\leq$ 2.7 pp for filler-gap; $\leq$ 0.3 pp for SV agreement), confirming that findings are not artefacts of any particular lexicon scrambling.

2.5 Evaluation

We evaluate using the Zorro benchmark (Huebner et al., 2021), a grammaticality judgement test suite designed for child-directed vocabulary. Zorro tests 13 phenomena across 23 paradigms using minimal pairs. We report accuracy (percentage of pairs where the model assigns higher probability to the grammatical sentence; ties count as incorrect) for each phenomenon and overall, averaging across paradigms within each phenomenon.

2.6 Experimental Design

We conduct 130 runs in total, with all hypotheses pre-registered on OSF after smoke testing confirmed implementation correctness and before the main experimental grid was executed. **Baselines** (5 runs): BabyBERTa trained on 100% data with 5 random seeds (42, 123, 456, 789, 1001). **Real grammar** (105 runs): 7 λ values (0.001, 0.01, 0.05, 0.1, 0.2, 0.5, 1.0) $\times$ 3 data fractions (25%, 50%, 100%) $\times$ 5 seeds. **Random grammar** (60 runs): 3 independent lexicon permutations (seeds 99, 2026, 31415) $\times$ 2 λ values (0.2, 0.5) $\times$ 2 data fractions (25%, 50%) $\times$ 5 seeds. Training steps are scaled by data fraction (25K / 50K / 100K). The λ range spans from negligible (0.001, where the grammar gradient is present but too weak to reduce grammar

loss) to strong (1.0, where the weighted grammar term contributes more to the total loss than the MLM term).

All other hyperparameters were kept at BabyBERTa defaults. Confirmatory analyses use Bonferroni correction ($k = 7$ λ values for H1, yielding per-test $\alpha = 0.05/7 = 0.007$; $k = 3$ pairwise comparisons for exploratory three-condition tests, yielding $\alpha = 0.05/3 = 0.017$); exploratory analyses (including the random grammar control) are labelled as such throughout.

3 Results

3.1 Overall Accuracy and Confirmatory Hypotheses

Our pre-registered primary hypothesis (H1) predicted that grammar constraints would improve overall Zorro accuracy. This hypothesis was **falsified**: the best constrained model at 100% data ($\lambda = 0.01$, 69.1%) did not significantly outperform the baseline (68.8%; $t(8) = 0.59$, $p = 0.574$, $d = 0.37$; Bonferroni-corrected $\alpha = 0.007$). Cohen's d throughout uses the pooled standard deviation across seeds. Overall accuracy decreased monotonically above $\lambda = 0.01$ (Figure 1). Our sample efficiency hypothesis (H2) was also falsified: the best constrained model at 50% data (65.1%) fell significantly short of the baseline at 100% data (68.8%; $t(8) = -6.22$, $p = 0.003$, $d = -3.84$). Our specificity hypothesis (H3) predicting greater improvement for grammar-encoded phenomena was not supported at any data fraction (100%: $p = 0.074$; 25%: $p = 0.08$; 50%: $p = 0.35$).

BabyBERTa is already a strong baseline for locally cued phenomena, with subject-verb agreement accuracy near 62% and case near 94%, leaving limited headroom for improvement on these tasks. However, these null overall results mask phenomenon-specific patterns that emerge from the random grammar control analysis.

3.2 Filler-Gap Dependencies: Linguistic Specificity

The random grammar control reveals that filler-gap accuracy gains are *linguistically specific*: real grammar outperforms random grammar by a substantial and statistically significant margin (Figure 2).

At 50% data with $\lambda = 0.5$: real grammar achieves 91.0% filler-gap accuracy versus 82.1% for random grammar (mean across three permutations) versus 79.2% for baseline. The real–random gap of +8.8 pp is highly significant ($t(18) = 5.47$, $p < 0.001$, $d = 2.82$; Mann–Whitney $U = 75$, $p = 0.001$).[2] At 25% data with $\lambda = 0.5$: the pattern is even more striking. Real grammar scores 82.2% versus random grammar at 69.3% (mean across three permutations) versus baseline at 79.2%, yielding a real–random gap of +12.9 pp ($t(18) = 4.67$, $p < 0.001$, $d = 2.41$; Mann–Whitney $p = 0.002$, 70/75 pairs). Notably, random grammar at this setting actually *decreases* filler-gap accuracy below baseline (−9.9 pp), while real grammar improves it (+3.0 pp).

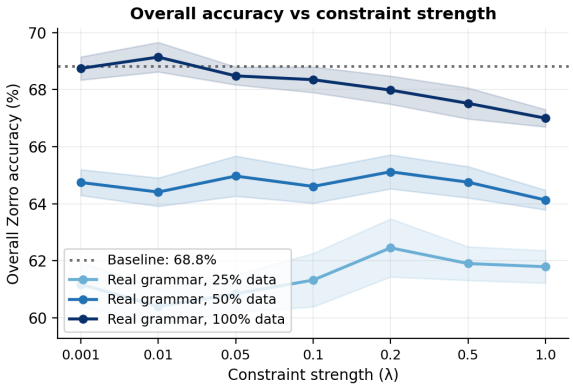


Figure 1: Overall Zorro accuracy vs. constraint strength λ for three data fractions. The dashed line shows baseline accuracy at 100% data (68.8%). Grammar constraints do not improve overall accuracy at any λ or data fraction, falsifying H1 and H2.

This three-condition comparison enables a principled decomposition of the grammar constraint's effect. The CKY computation itself provides some structural regularisation: random grammar at 50% data still outperforms baseline by ∼3 pp on filler-gap. However, accurate category assignments contribute substantially beyond this: approximately three-quarters of the total effect at 50% data (+8.8 pp linguistic out of +11.7 pp total, averaged across three permutations), and more than the entire effect at 25% data where random grammar actively hurts (−9.9 pp on average). As data becomes scarcer, the regularisation component diminishes while the linguistic specificity component becomes dominant.

Figure 2 plots filler-gap accuracy against λ

[2] All 75 real–random seed pairs show the real grammar exceeding random (75/75); bootstrap 95% CI for the gap: [5.4, 12.2] pp. Three independent lexicon permutations (seeds 99, 2026, 31415) yield consistent random means of 83.9%, 81.7%, and 80.8%, confirming the result is not permutation-specific. Bonferroni-corrected $\alpha = 0.017$ for three pairwise comparisons.

for both conditions. At low constraint strengths ($\lambda \leq 0.1$), real grammar performance is flat or slightly below baseline. Above this threshold, real grammar climbs steeply while random grammar diverges downward, visually confirming that the model specifically requires *correct* structural information to solve filler-gap dependencies, not merely any tree-shaped auxiliary loss.

3.3 Agreement, Binding, and NPI Patterns

Figure 3 shows the full pattern across focal phenomena for both grammar conditions.

Subject-verb agreement. Both real and random grammar reduce agreement accuracy by 8–11 pp below baseline in all conditions, and are statistically indistinguishable (all pairwise $p > 0.05$, $d \approx 0$). This symmetric damage rules out lexical interference as the cause; if misassigned categories were responsible, random grammar should damage agreement *more*. Instead, the CKY computation itself appears to divert representational capacity from the local lexical-distributional features (singular/plural morpheme correlations) that drive agreement, regardless of whether the lexicon is linguistically accurate.

NPI licensing. NPI licensing shows high instability across lexicon permutations. A single permutation (seed 99) suggested random grammar produced larger NPI gains than real grammar, but this does not replicate: across three permutations, random grammar averages 35.4% (barely above the 35.0% baseline) while real grammar averages 39.3%, with neither condition significantly differing ($t(18) = 0.66$, $p = 0.52$). The high per-seed variance (SD = 16.9; bimodal distribution)[3] indicates that NPI licensing is not robustly learnable from grammar constraints at this data scale, and underscores the value of multiple random controls.

Binding. Both grammars reduce binding accuracy below baseline at $\lambda = 0.2$ and 0.5. However, at very low constraint strengths ($\lambda = 0.001$) in our 100% data experiments, real grammar produces a substantial binding improvement (+6.2 pp) despite the grammar loss not decreasing during training. We return to this finding in §4.1.

[3]Per-seed NPI accuracy for real grammar at 50%, $\lambda = 0.5$: 20.7, 25.5, 37.5, 53.9, 58.9%.

Condition	Fill.	SVAgr	Bind.	NPI
Baseline (100%)	79.2±1.5	**61.5**±1.4	**66.8**±4.1	35.0±4.2
Real (50%, λ=0.5)	**91.0**±2.0	52.6±1.0	49.7±7.7	39.3±16.9†
Rand (50%, λ=0.5)	82.1±3.4	52.7±1.1	50.2±3.5	**35.4**±9.3
Real − Rand	+8.8*	−0.1	−0.5	+3.9

Table 1: Mean accuracy (%) ± SD on focal phenomena. Real grammar: 5 seeds. Random grammar: 15 runs across 3 independent lexicon permutations (seeds 99, 2026, 31415). *Significant: $t(18) = 5.47$, $p < 0.001$, $d = 2.82$; all 75 real–random seed pairs show real > random (87.9–93.5 vs. 75.1–86.6). †High variance reflects a bimodal distribution across seeds (20.7, 25.5, 37.5, 53.9, 58.9%); median = 37.5%. **Bold** = best per column. Fill. = Filler-gap; Bind. = Binding.

3.4 Training Dynamics

The grammar auxiliary loss does not impair language model training. Figure 4 shows MLM loss convergence for real and random grammar conditions at 50% data, $\lambda = 0.5$. The two conditions produce virtually identical MLM loss trajectories. Grammar loss itself, however, diverges between conditions: real grammar loss decreases during training, while random grammar loss plateaus after an initial drop, because the model cannot simultaneously satisfy the MLM objective and a grammar that assigns tokens to incorrect categories. This divergence provides direct evidence that the model extracts usable structural signal from the real grammar but not from the random grammar, even though both impose identical computational overhead.

3.5 Summary

Table 1 summarises the key three-condition comparisons for focal phenomena. The pattern is clear: accurate structural priors specifically benefit filler-gap dependencies (real > random > baseline) while both grammars equally damage agreement and produce comparable or reversed effects on other phenomena. The full 13-phenomenon breakdown is shown in Appendix A.

4 Discussion

4.1 Inductive Bias as Gradient Direction

An intriguing pattern emerges at very low constraint strengths. At $\lambda = 0.001$, the grammar loss remains high throughout training, meaning no explicit grammar learning occurs, yet binding accuracy improves by 6.2 pp over baseline. This dissociation between loss magnitude and behavioural effect is consistent with the PCFG objective acting

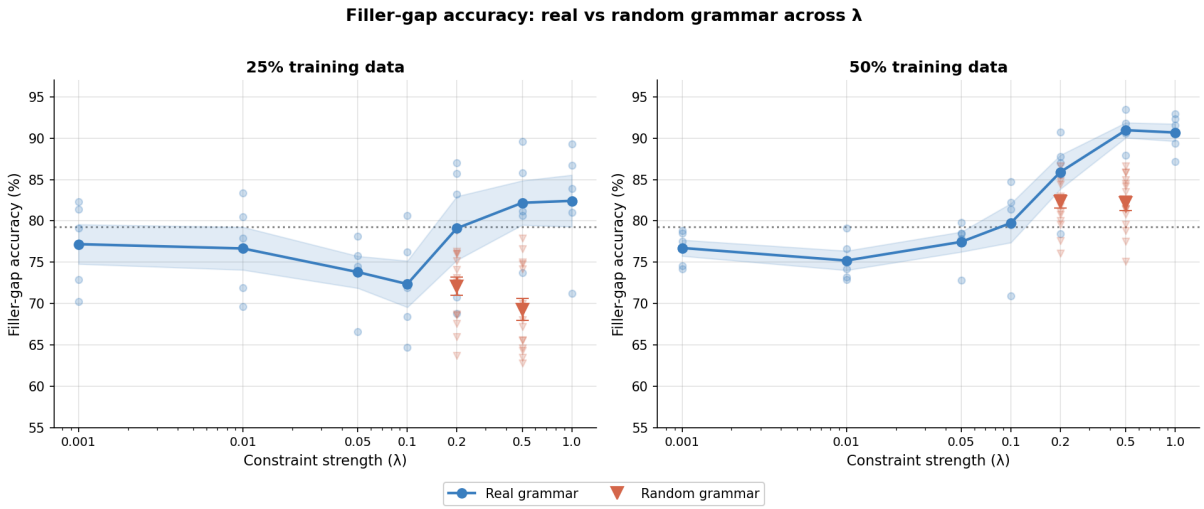


Figure 2: Filler-gap accuracy across the full λ sweep for real grammar (circles, solid) and random grammar (triangles, red) at 25% (left) and 50% (right) training data. At low λ, both conditions perform comparably. Above $\lambda \approx 0.1$, real grammar climbs steeply while random grammar diverges downward, particularly at 25% data. Baseline (100% data) shown as dashed line.

as a *directional bias* on the gradient rather than as an optimisation target: the grammar gradient may nudge the model's representations toward hierarchical structure even when too weak to measurably reduce grammar loss.

However, we note that this observation rests on a single λ value and a single phenomenon; we did not conduct the attention-pattern or probing analyses that would be needed to substantiate a mechanistic account. We flag it as a finding warranting dedicated investigation, particularly analysis of what changes in the model's representations at very low λ, rather than offering it as a confirmed explanation. Its consistency with accounts emphasising the role of inductive bias over explicit knowledge in acquisition (Griffiths et al., 2010; Tenenbaum et al., 2011; Yang, 2004) makes it theoretically interesting, but the empirical support here is preliminary.

4.2 Why Filler-Gap but Not Agreement?

The phenomenon-specific pattern has a natural linguistic interpretation rooted in the computational demands of each construction.

Filler-gap dependencies are inherently long-distance and hierarchical: they require tracking a displaced constituent across potentially unbounded clause boundaries (Wilcox et al., 2018, 2024; Howitt et al., 2024). This is precisely the kind of dependency where phrase-structure representations can encode genuine information about which constituents can be displaced and where gaps are licensed. Subject-verb agreement, by contrast, is heavily local and lexically cued; models can learn it through sequential distributional statistics without explicit hierarchical representations (Linzen et al., 2016; Gulordava et al., 2018). The CKY auxiliary loss at moderate-to-high λ values imposes a representational trade-off. By forcing the model to represent tokens as members of abstract phrase-structure categories, it diverts capacity from the fine-grained lexical features that drive agreement. This explains why the damage is symmetric across real and random grammar.

A further dissociation — between filler-gap and island constraints — sharpens this picture. Although both are encoded in the grammar and both involve long-distance structure, filler-gap accuracy rises from 79% to 91% at 50% data with $\lambda = 0.5$, while island accuracy *decreases* from 74% to 69% under the same condition. This suggests that structural priors help the model resolve dependencies (positive licensing) but not recognise where dependencies are prohibited (negative constraints). The syntactic interpretation is straightforward: a PCFG can encode the trace position of a moved element, licensing the filler-gap dependency itself, but island constraints require specifying where movement *cannot* occur, negative conditions not naturally expressed in phrase-structure grammar without feature-passing or constraint-based machinery.

The benefit of structural priors may therefore be limited to positive structural facts encodable in the grammar fragment, while negative constraints require richer representations. Notably, the grammar

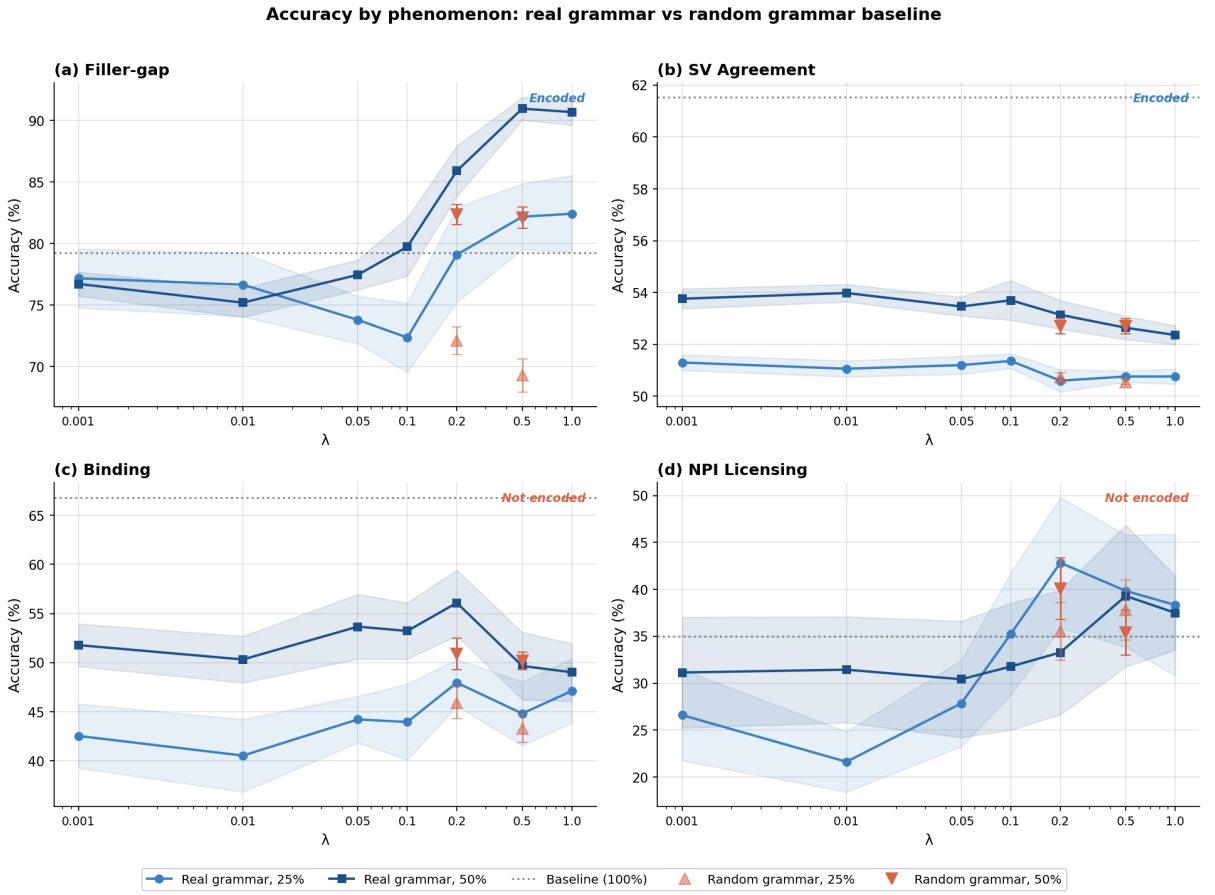


Figure 3: Accuracy by phenomenon: real grammar versus random grammar baseline across λ at 25% (circles) and 50% (squares) training data. Dashed baseline (100% data) shown for reference. (a) Filler-gap shows strong real-grammar advantage at high λ; (b) SV agreement shows symmetric damage from both grammars; (c) Binding improves with real grammar at very low λ despite no grammar loss decrease; (d) NPI licensing shows apparent random-grammar advantage, suggesting regularisation artefact.

loss is computed only on sentences truncated to 12 tokens, yet filler-gap improvements generalise to full-length Zorro items, suggesting that short-sentence structural supervision reshapes representations in ways that benefit long-distance processing.

4.3 Regularisation versus Linguistic Content

The three-permutation random grammar control enables a principled decomposition of observed effects. For filler-gap, the decomposition shifts across data scales: at 50% data, approximately three-quarters of the improvement comes from linguistically specific content (+8.8 pp) and one-quarter from structural regularisation alone (+2.9 pp above baseline); at 25% data, regularisation is not merely reduced but negative — random grammar hurts filler-gap by −9.9 pp — while linguistic content drives the entire benefit.

From a developmental perspective, this interaction has a natural interpretation as a learning-trajectory effect: under extreme input scarcity (25% data, ∼1.25M words), only linguistically accurate priors help — generic structural regularisation actively hurts. As input accumulates (50% data), regularisation begins to contribute positively, but accurate linguistic content still accounts for three-quarters of the benefit. This pattern suggests that the role of structural priors may shift during development; early acquisition, when data is most limited, would depend most critically on the accuracy of whatever structural biases are available.

These results also carry a practical warning. Single-permutation structural regularisation can produce unstable results on high-variance phenomena (as illustrated by the NPI finding in §3.3). Without multiple linguistically grounded control permutations, it would be easy to misattribute permutation-specific noise to principled linguistic content.

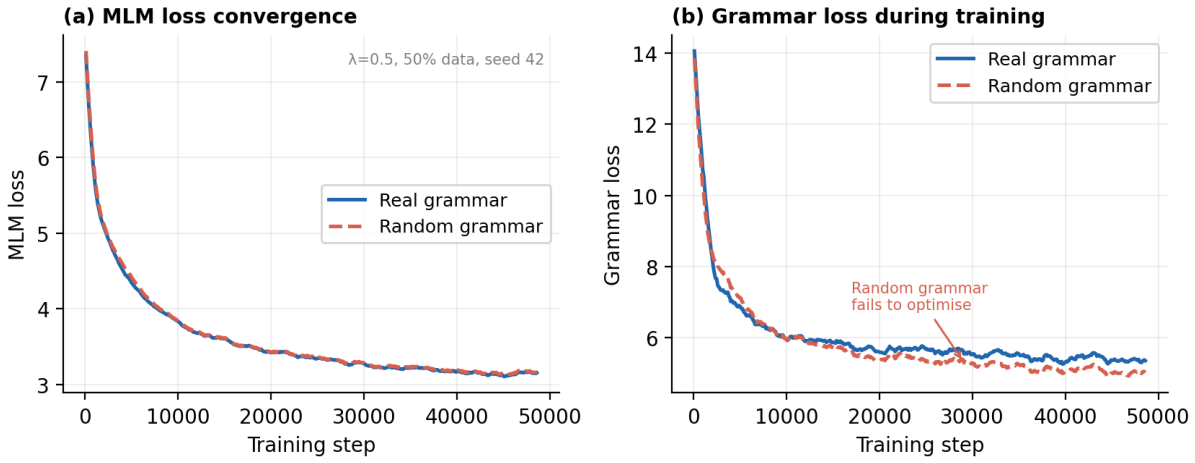


Figure 4: Training dynamics at 50% data, $\lambda = 0.5$. (a) MLM loss converges identically for real and random grammar. (b) Grammar loss decreases for real grammar but plateaus for random grammar, confirming that the model extracts usable structural signal only from the linguistically accurate lexicon.

4.4 Relation to Prior Work

Where the BabyLM tradition (Warstadt et al., 2023, 2024) asks *what* grammatical knowledge distributional learning can achieve, we ask *how* the learning objective changes the learning trajectory, treating grammar priors as an active intervention on the training signal. Within neurosymbolic grammar research (Kim et al., 2019; Yang et al., 2021; Park and Kim, 2025), we demonstrate that differentiable grammar integration provides targeted benefits at very small scales. Chesi et al. (2024) also integrate Minimalist Grammar constraints into BabyLM training, but by modifying RNN gating mechanisms to encode c-command and locality rather than adding a differentiable auxiliary loss to a transformer. Their eMG-RNN achieved comparable overall BLiMP scores to standard LSTMs but did not use a random grammar control, making it difficult to isolate the contribution of linguistic content from architectural regularisation.

Alternative routes for injecting structural bias include distilling from Bayesian models trained on formal languages (McCoy and Griffiths, 2025) and pre-pretraining on formal languages to impart linguistic biases (Hu et al., 2025). Our differentiable PCFG approach differs in providing continuous structural supervision during training rather than transferring structure from a separate pre-training phase. Whether these alternative routes produce the same phenomenon-specific pattern we observe is an open question.

Within the nativism/empiricism debate (Chomsky, 1965; Pullum and Scholz, 2002; Clark and Lappin, 2010; Chater and Manning, 2006), we provide empirical evidence that structural priors are *useful* for some aspects of grammar but not others. Filler-gap dependencies, which feature prominently in debates about the necessity of innate structural knowledge, are the constructions that benefit most from accurate structural supervision.

5 Conclusion

We tested whether explicit structural grammatical priors improve neural language model learning from child-scale data. Our pre-registered hypotheses about overall improvement and sample efficiency were falsified. However, a random grammar control with three independent lexicon permutations revealed that real grammar priors specifically improve filler-gap dependency learning beyond structural regularisation, with large effect sizes ($d = 2.41$–2.82). Wrong grammar damages filler-gap performance while failing to affect agreement. These phenomenon-specific effects suggest that the value of structural priors depends on the computational demands of the grammatical dependency: long-distance hierarchical constructions benefit from accurate linguistic knowledge, while locally cued phenomena do not.

Future work should test whether richer grammars encoding movement and feature-checking produce broader benefits, and whether these effects replicate cross-linguistically. It would also be valuable to investigate whether probing classifiers or attention-pattern analyses can identify the representational changes underlying the gradient-direction

effect at low λ, and whether the phenomenon-specific pattern persists in larger architectures, connecting the developmental "growing up" approach to the scalability concerns of modern language modelling.

More broadly, we propose the random grammar control as a methodological standard for neurosymbolic language model research. Without a linguistically scrambled baseline, apparent improvements from structural supervision cannot be distinguished from generic regularisation. Our results show that a single lexicon permutation may be insufficient; we recommend at least three independent permutations.

For developmental theory, our results suggest that the contribution of structural biases to language acquisition is neither uniform nor absent, but phenomenon-specific — with the strongest effects emerging precisely where input alone is most impoverished relative to the computational demands of the target construction.

Limitations

Several limitations qualify these results. First, our grammar is a simplified PCFG fragment, not a full Minimalist Grammar. Second, we evaluate only English child-directed speech. Third, all three confirmatory hypotheses (H1–H3) were falsified; the strongest findings come from the exploratory random grammar analysis. Fourth, the bimodal per-seed distribution for NPI licensing ($\mathrm{SD} = 16.9$) indicates that at child-scale data volumes, random initialisation can dominate the effect of grammar constraints for fragile phenomena. Future work should report full seed distributions rather than means alone. Fifth, our pre-registration also specified data-derived grammar and frequency-based control conditions, which were not implemented due to scope constraints; the exploratory developmental-trajectory (H4) and error-analysis (H5) hypotheses were likewise deferred to future work.

References

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

Nick Chater and Christopher D. Manning. 2006. Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, 10(7):335–344.

Cristiano Chesi, Veronica Bressan, Matilde Barbini, Achille Fusco, Maria Letizia Piccini Bianchessi, Sofia Neri, Sarah Rossi, and Tommaso Sgrizzi. 2024. Different ways to forget: Linguistic gates in recurrent neural networks. In *Proceedings of the 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 106–117.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Alexander Clark and Shalom Lappin. 2010. *Linguistic Nativism and the Poverty of the Stimulus*. Wiley-Blackwell.

Thomas L. Griffiths, Nick Chater, Charles Kemp, Amy Perfors, and Joshua B. Tenenbaum. 2010. Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8):357–364.

Kristina Gulordava, Piotr Bojanowski, Édouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL-HLT*, pages 1195–1205.

Kyle Howitt, Suraj Nair, Andrew Dods, and Robert M. Hopkins. 2024. Generalizations across filler-gap dependencies in neural language models. *Preprint*, arXiv:2410.18225.

Michael Y. Hu, Jackson Petty, Chuan Shi, William Merrill, and Tal Linzen. 2025. Between circuits and Chomsky: Pre-pretraining on formal languages imparts linguistic biases. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9691–9709.

Philip A. Huebner, Elior Sulem, Cynthia Fisher, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of CoNLL*, pages 624–646.

Yoon Kim, Chris Dyer, and Alexander Rush. 2019. Compound probabilistic context-free grammars for grammar induction. In *Proceedings of ACL*, pages 2369–2385.

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.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum.

R. Thomas McCoy and Thomas L. Griffiths. 2025. Modeling rapid language learning by distilling Bayesian priors into artificial neural networks. *Nature Communications*, 16(1):4676.

Jiyeon Park and Kangil Kim. 2025. Probability distribution collapse in unsupervised neural grammar induction. In *Proceedings of EMNLP*, pages 33392–33403.

Lisa Pearl. 2022. Poverty of the stimulus without tears. *Language Learning and Development*, 18(4):415–454.

Geoffrey K. Pullum and Barbara C. Scholz. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review*, 19(1–2):9–50.

Edward Stabler. 1997. Derivational minimalism. In Christian Retoré, editor, *Logical Aspects of Computational Linguistics*. Springer.

Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285.

Alex Warstadt and Samuel R. Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic Structures in Natural Language*. CRC Press.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjape, 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.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan G. Wilcox, Chengxu Zhuang, Adina Williams, Ryan Cotterell, and Tal Linzen. 2024. Insights from the first BabyLM Challenge: Training sample-efficient language models on a developmentally plausible corpus. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Ethan G. Wilcox, Richard Futrell, and Roger Levy. 2024. Using computational models to test syntactic learnability. *Linguistic Inquiry*, 55(4):805–848.

Ethan G. Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler-gap dependencies? In *Proceedings of BlackboxNLP*, pages 211–221.

Charles Yang. 2004. Universal Grammar, statistics, or both? *Trends in Cognitive Sciences*, 8(10):451–456.

Songlin Yang, Yanpeng Zhao, and Kewei Tu. 2021. PCFGs can do better: Inducing probabilistic context-free grammars with many symbols. In *Proceedings of NAACL-HLT*, pages 1487–1498.

A Full Phenomenon Heatmap

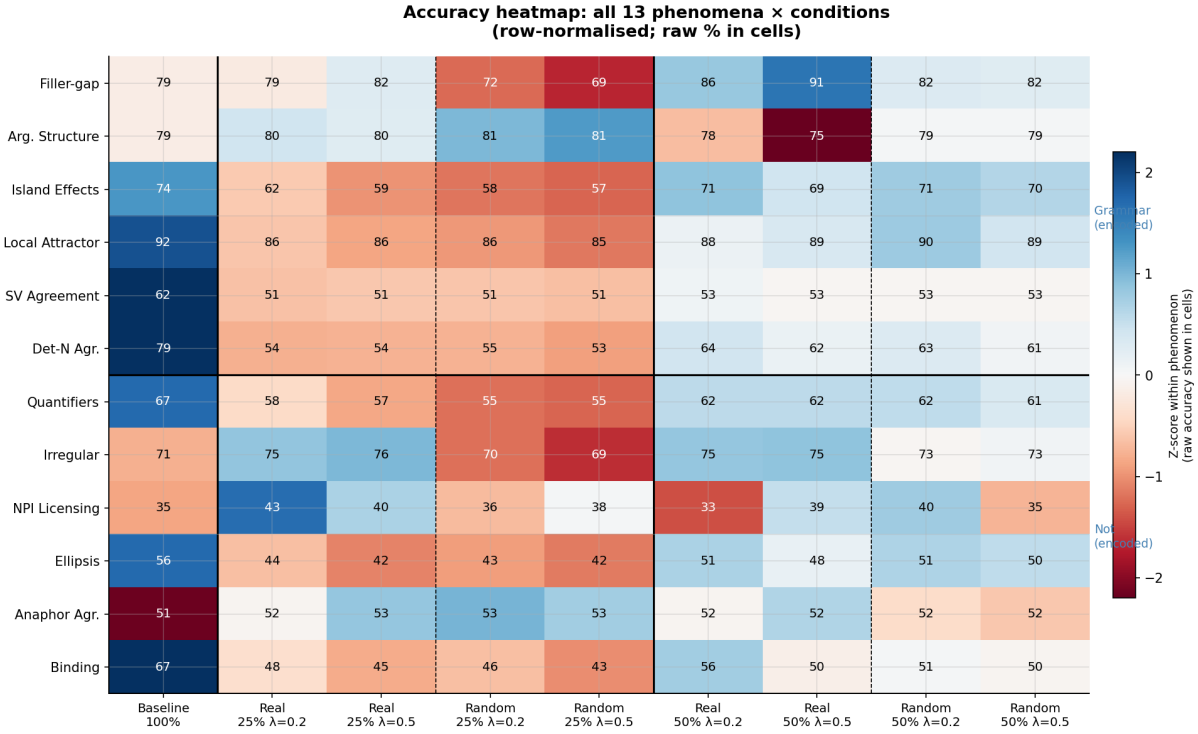


Figure 5: Accuracy heatmap across all 13 phenomena and 9 conditions (row-normalised; raw accuracy shown in cells). Conditions are organised by data fraction (25%, 50%) and constraint type (real / random), with the 100% baseline shown at left. Strong blue = high relative accuracy; strong red = low relative accuracy. The clearest cross-phenomenon contrast is between **Filler-gap** (row 1) — where Real 50%/λ=0.5 is the single darkest blue cell in the matrix (91%) — and **SV Agreement** (row 5), where all grammar-constrained conditions show uniform red regardless of whether the lexicon is real or random. This visual contrast directly reflects the paper's central finding: filler-gap improvement is linguistically specific, while SV damage is not. NPI Licensing (row 9) shows the highest condition-by-seed variance; binding and case show partial floor/ceiling effects at 25% data.

B Grammar Construction

The PCFG was hand-crafted to capture core English phrase structure as described by Minimalist Grammar (Stabler, 1997), implemented in Chomsky Normal Form for compatibility with the CKY inside algorithm. We describe its components, construction rationale, and limitations.

Production rules (16 rules). The grammar encodes basic X-bar structure: sentences are TPs dominating NP subjects and VP predicates (TP $\rightarrow$ NP VP). VPs decompose into verbs with complements (VP $\rightarrow$ V NP; VP $\rightarrow$ V CP). Embedded clauses use CP $\rightarrow$ C TP, which also serves as the filler-gap licensing configuration: the complementiser position marks where a wh-element can be base-generated, with the TP providing the clause containing the gap. Determiner phrases use DP $\rightarrow$ D NP. Coordination, adjunction, and PP attachment are included as binary-branching rules.

Terminal categories (12). N (noun), V (verb), D (determiner), Adj (adjective), Adv (adverb), P (preposition), C (complementiser), Conj (conjunction), Pro (pronoun), Aux (auxiliary), Neg (negation), Wh (wh-word). These were chosen to cover the major lexical and functional categories relevant to the Zorro phenomena.

Lexicon construction (7,230 entries). Each word type in AO-CHILDES was assigned to one or more grammatical categories based on its predominant usage in child-directed speech, informed by the CHILDES %mor tier annotations. Polysemous items (e.g., "run" as N or V; "that" as D, C, or Pro) receive multiple category assignments; the SoftLexicon routing distributes probability mass across all assigned categories proportionally. Approximately 12% of word types have multiple category assignments.

What the grammar encodes and does not encode. The grammar directly encodes hierarchical phrase structure, basic subcategorisation (transitive vs. intransitive verbs), determiner-noun cooccurrence, and the CP $\rightarrow$ C TP configuration that licenses filler-gap dependencies. It does *not* encode: feature-checking or Agree relations (so agreement is structurally represented but not enforced), binding domains (Principle A/B/C), negative polarity licensing contexts, island constraints (which require specifying where movement is prohibited), or quantifier scope. This limitation is by design: phenomena not encoded in the grammar serve as natural controls for specificity.

C Per-Seed Accuracy

Table 2 reports per-seed accuracy for focal phenomena at the key conditions discussed in the main text.

Condition	Seed	Fill.	SVAgr	Bind.	NPI
Baseline	42	79.8	60.2	61.7	29.1
	123	77.8	63.6	64.5	38.1
	456	77.7	60.1	71.9	33.6
	789	81.1	61.8	69.7	34.5
	1001	79.9	61.9	66.4	39.7
Real (50%, λ=0.5)	42	91.0	52.9	43.5	20.7
	123	90.6	52.8	62.6	25.5
	456	87.9	54.1	46.7	58.9
	789	93.5	51.5	50.4	53.9
	1001	91.8	51.9	45.1	37.5
Random, lex 99 (50%, λ=0.5)	42	86.6	54.1	54.0	37.5
	123	81.5	51.6	53.3	37.2
	456	84.2	53.8	52.3	49.9
	789	85.9	50.8	44.8	40.2
	1001	81.5	53.0	50.3	55.3
Random, lex 2026 (50%, λ=0.5)	42	85.9	53.4	51.3	27.3
	123	79.6	53.3	55.3	27.0
	456	80.8	53.3	50.2	40.7
	789	77.5	51.6	53.5	34.0
	1001	84.5	52.6	49.8	28.9
Random, lex 31415 (50%, λ=0.5)	42	81.4	53.7	51.1	24.4
	123	83.5	51.1	50.3	24.2
	456	85.0	54.0	47.5	35.2
	789	78.8	51.4	45.1	42.5
	1001	75.1	52.7	43.7	26.4

Table 2: Per-seed accuracy (%) on focal phenomena for key conditions. Real grammar filler-gap scores range 87.9–93.5%, with all 5 seeds above any of the 15 random grammar seeds at this condition. NPI licensing shows the bimodal pattern discussed in §3.3, with real grammar seeds splitting into low (20.7, 25.5) and high (37.5, 53.9, 58.9) clusters.

Self-Supervised Speech Representations Track Spoken Language Convergence to Adult Models in Infants and Children Who Are Deaf/Hard-of-Hearing

Landon Choy, Ali Sartaz Khan, Sonia Patrizi,
Daisy Ye, Julianna Gross, Margaret Cychosz
Stanford University, USA
mcychosz@stanford.edu

Abstract

Language development is characterized by a gradual convergence of children's speech toward adult patterns. Measuring this process has traditionally required detailed transcription and language-specific expertise, limiting scalability across languages and populations. Here, we use speech embeddings to capture this convergence directly from the acoustic signal in longform, child-centered recordings, taken as children go about their daily lives. Using HuBERT-BASE, we extracted embeddings from speech vocalizations of children who are deaf/hard-of-hearing and their female adult caregivers (>925 hrs. observation). Embedding distance between children and caregivers decreased with hearing age, controlling for pitch and vocalization length, indicating, as expected, that children's speech patterns converge to caregivers over development. This single distance metric likewise related to multiple standardized measures of speech and language from infancy through preschoolhood. These results suggest a path toward scalable, language-neutral assessment of spoken language development from children's everyday lives.

1 Introduction

Automatic recognition of children's spontaneous speech offers an opportunity to derive objective measures of language development from audio, forgoing the costly transcription and linguistic expertise traditionally required (Demuth et al., 2006). However, naturalistic, usually child-centered, recordings contain background noise, overlapping speech, and are plagued by diarization errors, making it difficult to derive meaningful language signals or developmental metrics (Li et al., 2025; Peurey et al., 2025). Developing methods that remain robust under these conditions is an open problem for automatic speech recognition research.

Recent work has begun to address this challenge. Sy et al. (2023) proposed an unsupervised metric of language development based on the entropy of discretized speech units derived from HuBERT-BASE embeddings (Hsu et al., 2021). Entropy quantified how surprising children's speech was under an adult-trained language model, with lower entropy indicating increasing convergence toward adult speech. While this approach failed under noisy naturalistic recordings, experiments with clean synthetic speech recovered the expected pattern of child entropy converging toward caregivers. Ott and Cychosz (to appear) showed that deriving canonical proportion, the proportion of well-formed consonant-vowel transitions, from naturalistic, child-centered audio predicts preschoolers' performance on numerous standardized measures of speech and language. But this measure is somewhat coarse, and may not capture finer-grained developments in children's speech or language development.

In this paper, we propose a simple framework that models language development as the continuous distance from caregiver speech in embedding space, without requiring transcription or discrete linguistic units. We apply this framework to children who are deaf or hard-of-hearing (DHH), who are at increased risk for language delay because their reduced access to spoken language can affect early language acquisition, particularly prior to intervention. Following intervention (e.g., hearing aids or cochlear implants), these children are under clinical care to ensure developmental progress, yet existing speech-language assessments are difficult to administer frequently and reliably, especially in infants and toddlers. This gap motivates scalable measures that can model speech-language development directly from children's everyday speech. The simplicity of our approach makes it computationally lightweight and practical for longitudinal follow-up of post-intervention progress.

We make the following contributions: 1) introduce an embedding-space distance metric com-

Proceedings of the 1st Workshop on Computational Developmental Linguistics (CDL), pages 27–36

puted directly from raw audio representations, capturing fine-grained developmental changes in children's speech-language structure relative to adult models; 2) show that this metric significantly predicts evaluated speech and language outcomes in children aged 10–65 months; and 3) demonstrate robustness to diarization errors through bootstrapping and sensitivity analyses, yielding stable estimates under input perturbations.

2 Methods

Participants were 34 children with bilateral moderate-profound hearing loss (16f/18m; 27 bilateral cochlear implant, 3 bimodal cochlear implant+hearing aid, 2 bilateral hearing aid, 2 unilateral cochlear implant); see Table 1 for age detail. 32/34 children were exposed to English >50%; two children were also exposed to some Mandarin (N=1) or Spanish (N=1). N=14 children (41%) contributed data from two or more longitudinal timepoints (M=2.8, SD=1.1, total observations in the dataset=59). Data collection was approved by the Institutional Review Boards at the authors' institution at the time of data collection.

Each child wore a Language ENvironment Analysis (LENA) recording device in a specialized shirt, capturing both surrounding speech and the child's own vocalizations. Families were instructed to activate the recorder once the child awoke and to record for up to 16 hours. Each recording, corresponding to a single child-timepoint, averaged 15.7 hours in length (SD = 1.3, range = 9.2–16), yielding a dataset of 925 hours of child-centered audio.

2.1 Speech-language assessments

Children/caregivers additionally completed a number of standardized speech-language assessments. *Parent-reported vocabulary*: Parents of children aged $\leq$ 52 mos. at study onset[1] completed the American English MacArthur-Bates Communicative Development Inventory (MB-CDI) (Fenson et al., 2007), a checklist of words the child knows and understands. Because we had recordings of the child's speech production in the long-form recordings, and not, for example, speech comprehension in a controlled task, here we model the number of words that children produced as one of our outcome measures in the results. *Child vocabulary* was measured in children $\geq$ 37 mos. at study onset with the Peabody Picture Vocabulary Test-4 (PPVT-4) (Dunn and Dunn, 2007), which indexes children's receptive vocabulary size, and the Expressive Vocabulary Test-2 (EVT-2) (Williams, 2007), which indexes expressive vocabulary size. *Child speech articulation* was assessed in children $\geq$ 37 mos. using the Goldman-Fristoe Test of Articulation-2 (GFTA-2) (Goldman and Fristoe, 2000), where consonant articulation accuracy is assessed across word positions (e.g. initial, medial) in a picture-naming task. Responses were scored offline by two trained research assistants. Speech-language assessments were conducted within 60 days (M=19.3, SD=22.0) of the child's longform recording to assess concurrent relationships between the spontaneous speech collected in the longform naturalistic recording and controlled speech-language patterns.

[1]Children who are DHH often have language delay; clinical judgment was applied to continue employing this assessment beyond the typical age range.

2.2 Processing pipeline

Recording speaker diarization was conducted by segmenting the continuous audio stream into speaker vocalizations using the LENA interpreted time segment '.its' (a proprietary data file format associated with each LENA recording) speaker labels. Audio segments classified as the key child (CHD; M=3530, SD=1346/rec) and adult female near the child (FEM; M=2391, SD=1141/rec) were retained (see Fig 1 for the full processing pipeline).

Embeddings were extracted from HuBERT-BASE, chosen because of its self-supervised pretraining mechanism that transfers well for downstream speech tasks across languages for both adult and child speech (Zanon Boito et al., 2024; Charlot et al., 2026). Embeddings were extracted from layers 7–9 following Charlot et al. (2026), who used layer 7 to generate pretraining clustering targets for downstream child voice type classification. Since the clustering procedure identifies hidden units corresponding to acoustic speech units, Charlot et al. (2026)'s pretraining procedure suggests HuBERT's mid-upper layers encode salient acoustic structure.

Fundamental frequency (f_0) was extracted using the PYIN algorithm (Mauch and Dixon, 2014). Here, f_0 is a control because speech representations encode f_0-related variation, particularly in lower transformer layers (Lin et al., 2023). Since f_0 decreases with anatomical growth over development, it could potentially confound linguistic convergence with acoustic maturation (Lee et al., 1999). Thus, we included each child's per-recording mean f_0 as a covariate to isolate linguistic from non-linguistic variance in embedding distance.

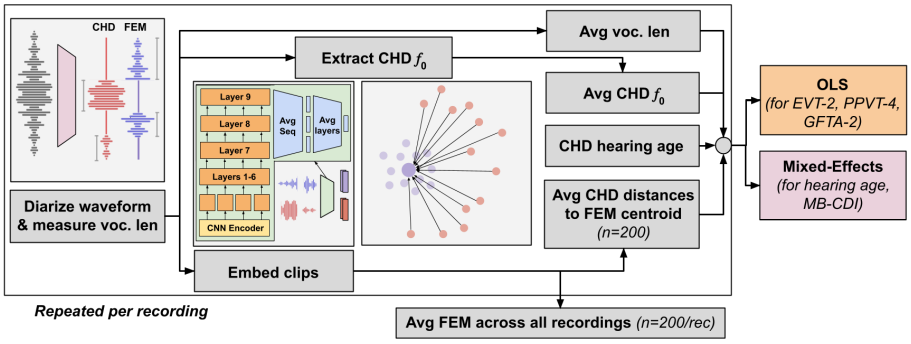


Figure 1: Processing pipeline overview. Longform recordings are speaker diarized (child 'CHD', adult female 'FEM') and embedded using HuBERT-BASE (Hsu et al., 2021). Audio was sampled at 16 kHz with a 20 ms frame rate, producing 768-dimensional vectors, which were extracted from layers 7–9 and mean-pooled across time and layers. Vocalization lengths were computed as the durations of speaker-diarized speech segments. Fundamental frequency (f_0) was extracted from subsampled key child (CHD) clips. Female adult (FEM) embeddings, representing adult female caregivers (e.g., mothers) from all recordings, were subsampled (small blue points) to create a global caregiver centroid (large blue point; visualized here as a 2D projection of the 768-dimensional embedding space). CHD embedding distance to this centroid was measured for each recording (red points) and used in statistical models as one observation/timepoint alongside mean f_0, mean vocalization length, and hearing age.

Variable	N_{obs}/N_{child}	M (SD)
Recording details		
Chron. age (mos)	59/34	36.2 (14.4)
Hearing age (mos)	59/34	22.2 (14.3)
Key child vocs/rec	59/34	3531 (1347)
Adult female vocs/rec	59/34	2392 (1142)
Duration (s)	59/34	16.1 (3.8)
Language outcomes		
MB-CDI	36/16	107.0 (166.3)
PPVT-4	22/17	95.6 (19.4)
EVT-2	23/18	100.1 (17.6)
GFTA-2	22/17	72.4 (17.3)

Table 1: Descriptive stats. N_{obs}=recording–assessment pairs; N_{child}=unique children; tp = timepoint. Hearing age=time since implantation/amplification. MB-CDI=raw words produced; PPVT-4, EVT-2, GFTA-2=standard scores.

Vocalization length was extracted from the .its files and likewise included as a covariate to avoid a potential confound with language development as children's speech utterances lengthen with age (Rice et al., 2010; Ramsdell-Hudock et al., 2018).

Embedding distance was computed by constructing a global FEM centroid across *all* recordings, from all adult females, and comparing individual CHD vocalizations from each recording to this reference. Distances were mean-pooled at the child–timepoint level to yield a single observation per pair (Fig. 1), to model each child's speech relative to a single, stable adult reference.

2.3 Sensitivity analysis

Speaker misattribution is a major concern in longform child-centered recordings. Following Gautheron et al. (2026), we conducted sensitivity analyses to simulate diarization errors and test whether the observed relationships could arise from speaker-label contamination. Specifically, increasing proportions of CHD vocalizations ($k\%$) were randomly replaced with vocalizations from other speaker classes ($k = 0$–100% in 10% increments). Embedding distances and downstream models were recomputed at each contamination level. Full details are provided in App. B.1.

3 Results

We begin by characterizing the embedding distance metric, and then evaluate its relationship with the children's hearing age to establish how it behaves given known language development trajectories. We then assess the relationship between embedding distance and standardized speech-language measures to test whether greater convergence to adult speech indexes more mature speech and language skills[2].

When the longitudinal structure of the data permitted (i.e. multiple observations per child), we fit linear mixed-effects models with a random intercept for child using `mixedlm` from `statsmodels`; otherwise, we fit ordinary least squares models using `ols` (Seabold and Perktold, 2010). Mean CHD f_0 from each recording was included as a covariate in all models to control for pitch-related variance in the embedding space (see Methods for

[2]Code for all analyses is available at: https://github.com/spoglab-stanford/cld-indexing

detail). Hearing age and vocalization length were also included to account for residual developmental variance not captured by f_0. Continuous predictors were z-score normalized. Model fit was evaluated using AIC and likelihood-ratio tests comparing models with and without the embedding distance term, alongside fixed-effect coefficient significance. We also report incremental variance explained (ΔR^2), the increase in R^2 attributable to each predictor when added to the model.

For each recording, 200 CHD vocalization embeddings/recording were sampled (from up to 800 observations) to compute the average distance to the FEM centroid. This subsampling was repeated for 1000 iterations, refitting the statistical models each time to obtain 95% bootstrap confidence intervals over the estimated coefficients and model metrics. The FEM centroid was fixed across iterations and computed from caregiver vocalizations pooled across all child-timepoints ($n = 59$), totaling 11,800 FEM vocalizations (200 per timepoint). The bootstrap intervals therefore capture variability in the sampled CHD observations and provide robustness to diarization errors.

3.1 Does embedding distance decrease with increased hearing experience?

We first tested the relationship between embedding distance and children's hearing age, or their experience with spoken language. We expect a gradual convergence of child to adult speech in the embedding space as children gain more experience. We found embedding distance improved baseline model fit and significantly predicted hearing age ($\beta = -0.50$, $p < .001$; Table 2; Fig 2a). Model comparisons further supported including embedding distance (ΔAIC$= -24.68$, $p < .001$), and it accounted for substantial additional variance (11.30%, $p < .001$). This indicates that hearing age, reflecting cumulative experience with spoken language, is associated with reduced child-adult embedding distance, suggesting progressive convergence toward adult speech patterns with increasing auditory experience. Hearing age was used throughout all analyses, rather than chronological age, as it provided better model fit (see App. A).

3.2 Does embedding distance index concurrent vocabulary size?

We next examined the relationship between embedding distance and children's vocabulary where we hypothesized that as CHD–FEM embedding distance decreased, vocabulary size would increase, which was consistent across all vocabulary measures (Figs 2b–d): embedding distance significantly predicted MB-CDI ($\beta = -0.35$, $p < .001$), PPVT-4 ($\beta = -0.29$, $p < .001$), and EVT-2 ($\beta = -0.24$, $p < .001$), and scores (Table 2). Model comparisons showed improved fit over baseline when including distance for MB-CDI (e.g., ΔAIC $= -10.28$, $p < .001$), with distance explaining additional variance for all outcomes (3.51–9.47%, $p < .001$). Although embedding distance remained a significant predictor of PPVT-4 and EVT-2 scores and explained additional variance, it did not significantly improve model fit. This may indicate that embedding distance captures developmental variance related to receptive and expressive vocabulary, but partially overlaps with age and acoustic covariates, limiting its unique contribution to model fit.

3.3 Does embedding distance index concurrent consonant articulation skill?

Our third analysis evaluated whether embedding distance was related to children's consonant articulation ability, measured using the GFTA-2. We expected that smaller CHD–FEM embedding distances would correspond to more accurate consonant articulation, which was confirmed ($\beta = -0.41$, $p < .001$; Table 2; Fig 2e), indicating, as expected, that children with more mature, accurate consonant articulation skill are closer in embedding space to global adult speech models. Similar to PPVT-4 and EVT-2 experiments, including distance did not significantly improve overall model fit, but explained substantial additional variance in GFTA-2 scores (14.33%, $p < .001$), suggesting that embedding distance captured articulation-related developmental variation that partially overlapped with hearing age, f_0, and voc. length.

3.4 Are the results robust to diarization errors?

To assess the robustness of our proposed metric to upstream diarization errors, we conducted sensitivity analyses, simulating increasing levels of speaker misattribution (full results in Appendix B.2). Across outcomes, the standardized embedding-distance coefficient remained directionally stable, with smaller CHD–FEM embedding distance continuing to predict more mature developmental and language outcomes despite progressive attenuation with increasing contamination (App. B.2 Fig 3). For hearing age and MB-CDI,

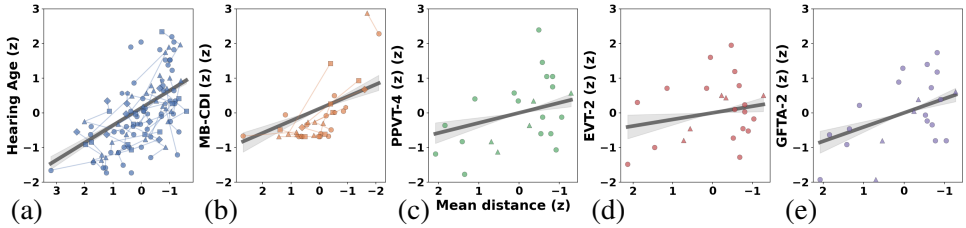


Figure 2: Relationships between CHD–FEM embedding distance and developmental measures. Points show child–timepoint mean distance from CHD embeddings to the FEM centroid, averaged across sampled clips (and bootstrap resamples). Marker shape indicates within-child timepoint order: first = circle, second = triangle, third = square, fourth = diamond, fifth = pentagon, and sixth = hexagon. Lines connect successive observations for the same child; connecting lines omitted for PPVT-4, EVT-2, and GFTA-2 due to limited repeated observations per child. The distance axis is reversed so that reduced CHD–FEM distance appears to the right, such that all panels show a negative trend. The gray line shows the fixed-effect relationship between distance and the outcome, computed from the mean bootstrap model coefficients; shaded ribbon indicates the pointwise 95% confidence interval obtained from the distribution of bootstrap-fitted lines.

Statistic	**Hearing Age**	**MB-CDI**	**PPVT-4**†	**EVT-2**†	**GFTA-2**†
Fixed-effect coefficients (with distance)					
Distance (β)	-0.50***	-0.35***	-0.29***	-0.19*	-0.41***
Hearing age (β)		0.53***	0.45***	0.34***	0.34***
Mean f0 (β)	-0.01	-0.09	-0.28***	-0.34***	-0.10
Mean voc. len (β)	0.01	0.12***	0.08	0.08	<0.001
Model comparison (with vs. without distance)					
ΔAIC	-24.68***	-10.28***	-0.84	0.79	-2.74
LR test $\chi^2(1)$	26.68***	12.28***	2.84***	1.21*	4.74***
LR p	<0.001***	<0.001***	0.13	0.35	0.05
Incremental variance explained (ΔR^2; beyond baseline predictors)					
Distance (%)	11.30***	9.47***	6.95***	3.51**	14.33***
Hearing age (%)		11.11***	18.22***	10.38***	10.39***

Table 2: Bootstrap model summaries. Coefficients correspond to the full model (with distance metric). ΔAIC and LR test compare models with vs. without distance, such that negative ΔAIC values favor the inclusion of our metric. (ΔR^2)=drop-one variance change in model R^2 attributable to each predictor beyond the others. Values show bootstrap means across CHD resamples; significance markers denote bootstrap significance (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). Bootstrap inference reflects CHD resampling, whereas LR test χ^2 statistics reflect model-level comparisons. †=OLS; others=mixed-effects with random intercept/child.

embedding distance continued to improve model fit relative to the baseline model across most contamination levels, whereas this pattern was not observed for PPVT-4, EVT-2, and GFTA-2, consistent with the original analyses in which distance effects were nonsignificant (App. B.2 Fig 4). The incremental variance remained directionally stable with progressing contamination like the embedding-distance coefficient (App. B.2 Fig 5). All outcomes exhibiting initial significance remained relatively robust under moderate contamination, suggesting the observed effects are unlikely to arise solely from systematic LENA speaker-label errors.

4 Conclusion

Child language research has traditionally relied on labor-intensive manual transcription, limiting scalability; consequently, only ~103 of the world's 7,000+ languages are represented in major language acquisition journals (Kidd and Garcia, 2022). We propose a simple alternative approach: measuring acoustic distance between children and caregivers within the same community directly from speech embeddings. We found that this distance decreases over development—reflecting convergence to adult speech—and is associated with speech-language outcomes. Critically, across models, developmental periods, and measures, embedding distance remained a significant predictor even after controlling for child age, f_0, and vocalization length, suggesting embedding distance captures variance beyond anatomical maturation. We further show that these results are robust to diarization errors through sensitivity analyses, suggesting the results are not artifacts of speaker diarization error. Together, these results suggest a simple and scalable technique to measure children's language development from their everyday speech patterns.

Limitations

We emphasize that embedding distance is an alternative measure of child language intended to complement, rather than replace, clinical analysis and careful characterization of child language patterns by trained professionals. While promising, the measure does not yet capture specific aspects of language development (e.g., morphological productivity), and its diagnostic utility remains unestablished. There is no evidence that it can be used diagnostically. Furthermore, an acoustic measure based on the child's own speech production will only indirectly index the child's receptive capabilities (e.g., ability to distinguish between phonological categories). Although we controlled for mean f_0 and vocalization length to account for developmental changes in vocal anatomy and utterance duration, embedding distance may still capture additional acoustic or non-linguistic sources of variation that require further investigation. This metric could also be extended to larger cohorts of children with typical hearing and controlled receptive outcomes, including speech perception tasks, to further evaluate its relationship to receptive language development.

Ethical considerations

The model employed, HuBERT-BASE, which was pretrained on 960 hours of English *LibriSpeech* audiobook speech (Hsu et al., 2021; Panayotov et al., 2015); therefore, the learned representations reflect monolingual English pretraining rather than multilingual exposure. The children reported in this work were primarily acquiring American English. It will thus be critical, going forward, to evaluate the performance of the embedding distance metric, derived from models such as HuBERT, in additional languages that are under-represented in the foundation model's training data, as well as how this approach of embedding distance extends to children acquiring different languages, or combinations of languages.

Acknowledgments

The authors thank the families who participated in this research, as well as Amy Martinez. Additional thanks to Jan Edwards, Ben Munson, and Mary Beckman for generously sharing their data, portions of which were reused for this project; data collection for those data was originally funded by National Institute on Deafness and Other Communication Disorders grant R01DC02932. Additional data collection and compute resources were funded by a Hearing Health Foundation Emerging Research Grant to M.C.

References

Théo Charlot, Tarek Kunze, Maxime Poli, Alejandrina Cristia, Emmanuel Dupoux, and Marvin Lavechin. 2026. Babyhubert: Multilingual self-supervised learning for segmenting speakers in child-centered long-form recordings. *Preprint*, arXiv:2509.15001.

Alejandrina Cristia, Marvin Lavechin, Camila Scaff, Melanie Soderstrom, Caroline Rowland, Okko Räsänen, John Bunce, and Elika Bergelson. 2021. A thorough evaluation of the language environment analysis (lena) system. *Behavior Research Methods*, 53(2):467–486.

Katherine Demuth, Jennifer Culbertson, and Jennifer Alter. 2006. Word-minimality, epenthesis and coda licensing in the early acquisition of english. *Language and Speech*, 49(2):137–173.

Lloyd M. Dunn and Douglas M. Dunn. 2007. *PPVT-4: Peabody Picture Vocabulary Test*, 4th edition. Pearson Assessments, Minneapolis, MN.

Larry Fenson, Virginia A. Marchman, Donna J. Thal, Philip S. Dale, J. Steven Reznick, and Elizabeth Bates. 2007. *MacArthur-Bates Communicative Development Inventories: User's Guide and Technical Manual*, 2 edition. Paul H. Brookes Publishing Co., Baltimore, MD.

Lucas Gautheron, Evan Kidd, Anton Malko, Marvin Lavechin, and Alejandrina Cristia. 2026. Classification errors distort findings in automated speech processing: examples and solutions from child-development research. *Preprint*, arXiv:2508.15637.

Ronald Goldman and Macalyne Fristoe. 2000. *GFTA-2: Goldman-Fristoe Test of Articulation 2*, 2nd edition. Pearson Assessments, Minneapolis, MN.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Evan Kidd and Rowena Garcia. 2022. How diverse is child language acquisition research? *First Language*, 42(6):703–735.

S. Lee, A. Potamianos, and S. Narayanan. 1999. Acoustics of children's speech: developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, 105(3):1455–1468.

Jialu Li, Marvin Lavechin, Xulin Fan, Nancy L. McElwain, Alejandrina Cristia, Paola Garcia-Perera, and

Mark A. Hasegawa-Johnson. 2025. Automated analysis of naturalistic recordings in early childhood: Applications, challenges, and opportunities. *IEEE Signal Processing Magazine*, 42(6):16–34.

Guan-Ting Lin, Chi-Luen Feng, Wei-Ping Huang, Yuan Tseng, Tzu-Han Lin, Chen-An Li, Hung-yi Lee, and Nigel G. Ward. 2023. On the utility of self-supervised models for prosody-related tasks. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1104–1111.

Matthias Mauch and Simon Dixon. 2014. Pyin: A fundamental frequency estimator using probabilistic threshold distributions. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 659–663.

Carissa Ott and Margaret Cychosz. to appear. Connecting preschoolers' spontaneous speech patterns to future language skills: A three-year concurrent and longitudinal cohort study of canonical proportion as a developmental index. *Journal of Speech, Language, and Hearing Research*. To appear.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Loann Peurey, Marvin Lavechin, Tarek Kunze, Manel Khentout, Lucas Gautheron, Emmanuel Dupoux, and Alejandrina Cristia. 2025. Fifteen Years of Child-Centered Long-Form Recordings: Promises, Resources, and Remaining Challenges to Validity. In *Interspeech 2025*, pages 3948–3952.

Heather L. Ramsdell-Hudock, Andrew Stuart, and Douglas F. Parham. 2018. Utterance duration as it relates to communicative variables in infant vocal development. *Journal of Speech, Language, and Hearing Research*, 61(2):246–256.

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

Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and statistical modeling with python. *SciPy 2010*.

Yaya Sy, William N. Havard, Marvin Lavechin, Emmanuel Dupoux, and Alejandrina Cristia. 2023. Measuring Language Development From Child-centered Recordings. In *Interspeech 2023*, pages 4618–4622.

Kathleen T. Williams. 2007. *EVT-2: Expressive Vocabulary Test*, 2nd edition. Pearson Assessments, Minneapolis, MN.

Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. 2024. mHuBERT-147: A Compact Multilingual HuBERT Model. In *Interspeech 2024*, pages 3939–3943.

A Chronological age vs. Hearing age

A.1 Methods

We evaluated model fit using hearing age, defined as months since intervention, and chronological age, defined as months old, and compared their relative ability to explain developmental variance using model fit (ΔAIC) and incremental variance explained (ΔR^2). All models included the baseline predictors of f_0, vocalization length, and CHD–FEM embedding distance. They additionally included either chronological age or hearing age, but not both. Because the chronological age and hearing age models are not nested, likelihood ratio tests were not appropriate and were therefore not evaluated. Instead, model comparisons focused on ΔAIC, computed as the AIC of the chronological age model minus the AIC of the hearing age model, such that positive values indicate better model fit using hearing age instead of chronological age. We were not interested in the direction or magnitude of the age coefficients themselves, but rather in whether chronological age or hearing age provided a better overall fit to developmental outcomes beyond the mentioned baseline predictors.

A.2 Results

Hearing age generally provided a better account of developmental outcomes than chronological age (Table 3). Model comparisons favored hearing age for PPVT-4, EVT-2, and GFTA-2, with lower AIC values relative to chronological age (ΔAIC = 4.92, 3.13, and 2.95, respectively; all $p < .001$), indicating improved model fit. Consistent with this, hearing age explained substantially more additional variance than chronological age for PPVT-4 (19.19% vs. 6.38%), EVT-2 (11.06% vs. 1.34%), and GFTA-2 (11.14% vs. 2.30%). In contrast, MB-CDI showed little difference between age measures: model comparison slightly favored chronological age (ΔAIC = 1.59), and both predictors explained comparable additional variance (13.14% vs. 12.59%). These results suggest that hearing age more closely tracks later vocabulary and articulation outcomes, whereas early vocabulary measured by the MB-CDI is similarly captured by either age metric, hence our choice for using hearing age in the main results.

Statistic	MB-CDI	PPVT-4†	EVT-2†	GFTA-2†
Model comparison (chron. age vs. hearing age)				
ΔAIC	1.59	4.92***	3.13***	2.95***
Incremental variance explained (ΔR^2; beyond baseline and distance predictors)				
Chronological age (%)	13.14***	6.38***	1.34**	2.30**
Hearing age (%)	12.59***	19.19***	11.06***	11.14***

Table 3: Bootstrap model summaries. Convention follows Section 3 Table 2.

B Sensitivity analysis

B.1 Methods

A limitation of the LENA diarization algorithm is potential speaker-label error and misclassification (e.g. child vocalizations labeled as female adult), which may propagate into downstream analyses (Cristia et al., 2021; Gautheron et al., 2026). To assess the robustness of our findings to such errors, we conducted a sensitivity analysis in which we assumed that ($k\%$) of the subsampled key child (CHD) diarized clips were misclassified. To simulate label contamination, we adopted a simplifying assumption that the existing LENA speaker assignments were correct and, at each subsampling iteration, replaced ($k\%$) of the CHD clips with clips uniformly sampled from the female adult (FEM), male adult (MAL), and other child (OCHD) speaker categories. Importantly, FEM clips introduced during contamination were sampled from a disjoint pool and were excluded from the FEM centroid construction to avoid circularity between contamination samples and the adult reference representation. The resulting contaminated sample sets were processed through the identical statistical modeling and confidence-interval estimation pipeline used in the primary analyses to quantify the robustness of the CHD–FEM embedding distance for estimating hearing age and language outcomes under increasing levels of speaker-label noise.

We did not introduce contamination into the FEM centroid itself. Unlike the CHD sample pool, the FEM representation was constructed as an aggregated centroid across recordings, yielding a single reference point expected to be comparatively stable to individual diarization errors. Contamination was restricted to the CHD samples, reflecting the primary source through which speaker-label noise would influence the child–adult distance.

Analysis was performed as a contamination sweep from 0% to 100% in 10% increments.

B.2 Results

Fig. 3 shows how the standardized distance coefficient remained directionally stable across contamination levels. For hearing age and all language outcomes (MB-CDI, PPVT-4, EVT-2, GFTA-2), the estimated effect consistently remained negative, indicating that greater distance from the adult reference distribution continued to predict poorer developmental and language outcomes. Although the magnitude of the effect attenuated progressively with increasing contamination, the direction of the relationship did not reverse, even under severe perturbation levels ($\geq$ 80% contamination). As expected, statistical significance diminished as contamination increased and the signal-to-noise ratio decreased.

In Fig. 4, across hearing age and MB-CDI, the distance-augmented model (Model 2) consistently outperformed the baseline model (Model 1), as indicated by negative ΔAIC values. Although this advantage diminished with increasing contamination, Model 2 remained preferred across most contamination levels, suggesting that embedding distance provides incremental explanatory value beyond the baseline covariates even under substantial label corruption. In contrast, this pattern was not observed for PPVT-4, EVT-2, and GFTA-2. For these outcomes, the distance effect was already statistically nonsignificant in the original bootstrap analyses, and consequently the contamination sweep did not show a consistent advantage of Model 2 over the baseline model. Thus, the contamination analysis primarily demonstrates robustness for outcomes in which embedding distance exhibited an initial significant association.

In Fig. 5, the incremental variance explained by embedding distance generally decreased with increasing contamination, consistent with progressive degradation of the underlying developmental signal. Across outcomes, this manifested either as a monotonic reduction in the estimated variance explained or as widening confidence intervals that increasingly encompassed the null value. Never-

theless, embedding distance continued to account for a non-trivial proportion of variance across several outcomes, in some cases remaining substantial even under high contamination levels.

In sum, for outcomes in which the embedding-distance metric exhibited an initial statistically significant association, the estimated effects remained relatively stable under increasing contamination. The progressive attenuation of statistical relationships at higher contamination levels suggests that the observed effects are unlikely to arise solely from systematic speaker-label errors in the LENA diarization pipeline.

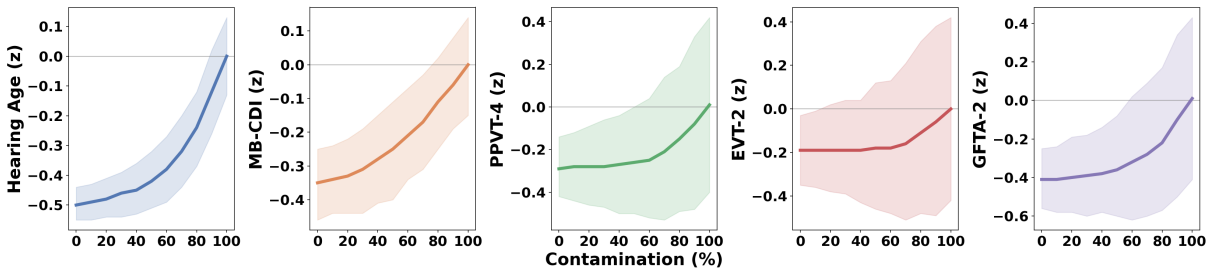


Figure 3: Standardized embedding-distance coefficient (β) under simulated speaker-label contamination. Contamination was swept from 0% to 100% in 10% increments. Solid lines show the mean estimated coefficient across bootstrap iterations, and shaded regions denote the corresponding 95% bootstrap confidence intervals.

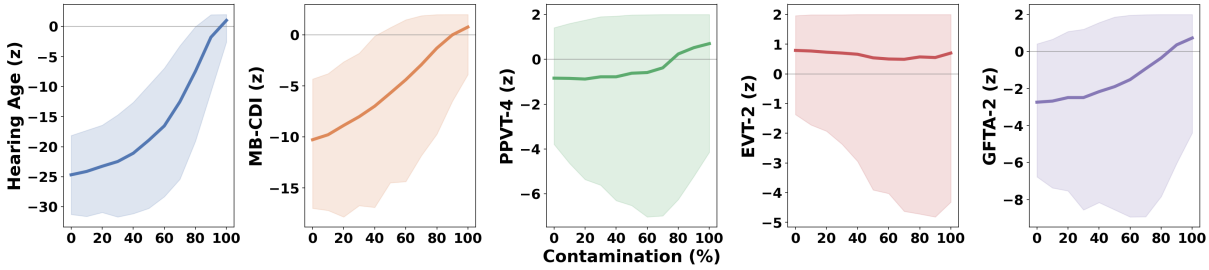


Figure 4: Model comparison metric (ΔAIC) under simulated speaker-label contamination. Convention follows Figure 3.

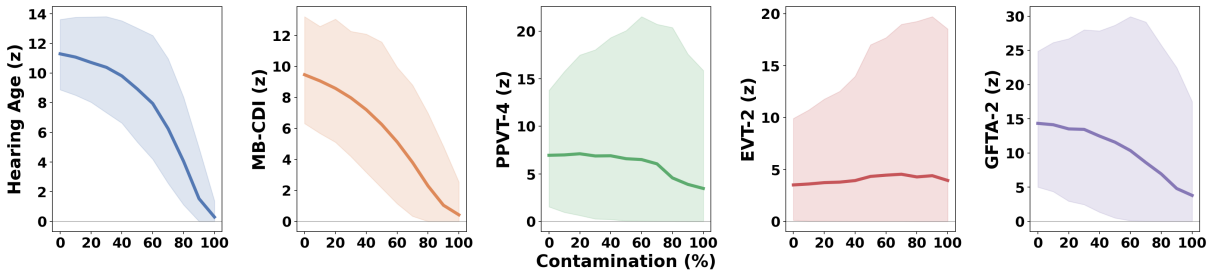


Figure 5: Incremental variance explained by embedding distance (% variance explained) under simulated speaker-label contamination. Convention follows Figure 3.

Do Language Models Show Structural Priming Across Different Domains?

So Young Lee[†], **Russell Scheinberg**[◇], **Ameeta Agrawal**[◇]
[†]Miami University, USA
[◇]Portland State University, USA
soyoung.lee@miamioh.edu
{rschein2,ameeta}@pdx.edu

Abstract

We test whether large language models show cross-domain structural priming by asking whether arithmetic expressions influence relative-clause attachment preferences. Experiment 1 examines English and French using materials based on prior psycholinguistic studies, and Experiment 2 extends the test to a larger multilingual dataset. Across both experiments, we find no robust priming effect. Instead, responses largely reflect baseline attachment preferences, which vary across languages and only partially align with human patterns. These findings suggest that, although language models show some structural sensitivity, they provide limited evidence of abstract structural generalization across domains.

1 Introduction

A central question in both cognitive science and natural language processing is whether the mechanisms that support language are domain-specific or domain-general. Research on human sentence processing has shown that structural priming—the tendency for prior exposure to a particular structure to bias subsequent interpretation—is not restricted to linguistic input alone, but can also be induced by structures from other domains, such as mathematics, logic, or music. These cross-domain effects suggest that human comprehenders recruit domain-general resources for representing and aligning hierarchical structures.

Recent NLP work increasingly treats language models as psycholinguistic test subjects and partial computational models of human sentence processing (Futrell et al., 2019; Wilcox et al., 2023; Cai et al., 2024).

Because LLMs are trained primarily on linguistic input, it remains unclear whether they exhibit priming effects that extend beyond within-language contexts. LLMs are trained on large text corpora

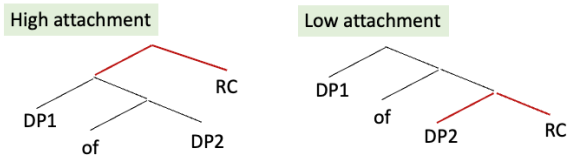


Figure 1: Syntactic Structures of DP1 of DP2 Modification (left) and DP2 Modification (right) in English

that include natural language, code and mathematical expressions, and they show substantial syntactic sensitivity within language, including long-distance agreement and other structure-sensitive contrasts (Linzen et al., 2016; Gulordava et al., 2018; Marvin and Linzen, 2018a; Warstadt et al., 2020). However, it remains unclear whether their structural representations are shared across domains in a way that supports cross-domain priming.

If models do show cross-domain priming, this would indicate that they encode abstract structural regularities that generalize across representational domains. If not, this would highlight a key boundary between human cognition and statistical language modeling. To investigate this issue, we focus on one of the most widely studied cases of syntactic ambiguity in human sentence processing: **attachment ambiguities**. For example, in (1), the relative clause (*who was on the balcony*) may attach either to the lower determiner phrase, DP2 (*the colonel*; low attachment), or to the higher determiner phrase, DP1 of DP2 (*the daughter of the colonel*; high attachment).

(1) The journalist interviewed the daughter of the colonel who was on the balcony.

Previous research has reported that English speakers tend to prefer low attachment. However, attachment preferences are not fixed: they can be modulated by prosody, lexical biases, discourse context, and, crucially, structural priming.

Cross-domain priming in sentence processing has been demonstrated in a series of psycholin-

Proceedings of the 1st Workshop on Computational Developmental Linguistics (CDL), pages 37–51

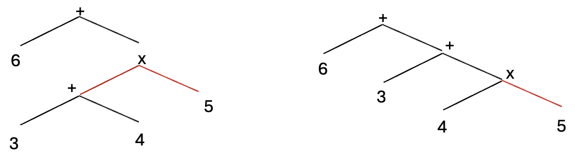


Figure 2: Hierarchical Structures of Arithmetic Expressions

guistic experiments. For instance, solving arithmetic expressions with nested groupings (e.g., $6 + (3 + 4) \times 5$) increases the likelihood of high relative clause attachment, whereas more linear expressions (e.g., $6+3+4\times5$) bias comprehenders toward low attachment (see Figure 2). These findings suggest that the mechanisms guiding syntactic disambiguation are sensitive to structural alignments across distinct representational domains. Rather than being tied exclusively to language, priming effects appear to reflect broader cognitive strategies for encoding and reusing hierarchical patterns.

LLMs, by contrast, have been observed to display priming-like behaviors only in linguistic contexts, such as persisting in syntactic choices across prompts (Prasad et al., 2019; Sinclair et al., 2022a; Michaelov et al., 2023; Jumelet et al., 2024a). Whether these effects extend to cross-domain contexts remains an open question. Because LLMs learn solely from textual corpora, they may not engage the same domain-general resources that humans recruit when transferring structural biases across domains. Testing whether LLMs exhibit cross-domain priming therefore provides a novel diagnostic of the extent to which their internal representations capture abstract structural parallels, or whether their priming behavior is confined to within-language statistical patterns. From a developmental perspective, the issue is not only whether language models eventually exhibit structure-sensitive behavior, but also what kind of representational organization emerges from their training experience. Human learners develop linguistic and non-linguistic reasoning abilities over time, and cross-domain priming has been taken as evidence that some aspects of hierarchical structure may be represented in a domain-general format. Language models provide a different kind of learner: they acquire behavior through large-scale exposure to text, code, mathematical expressions, and instruction-tuning. Testing whether arithmetic structure influences linguistic attachment therefore offers a way to ask whether model development gives rise to abstract, transferable structural representations, or whether linguistic and mathematical competencies remain functionally separated despite co-occurring in the training distribution. In this paper, we investigate this question by comparing human and LLM behavior in the same paradigm. Specifically, we ask whether models that have acquired both linguistic and mathematical competence show evidence of a shared structural representation across these domains. Building on prior psycholinguistic findings that mathematical grouping structures bias relative-clause attachment, we test whether contemporary LLMs show similar shifts. This comparison helps clarify the nature of priming, the extent of structural abstraction in LLMs, and the developmental trajectory through which such abstraction may or may not emerge in artificial learners.

2 Related Work

2.1 Structural Priming in Human Sentence Processing

Speakers tend to repeat syntactic structures they have recently encountered, a phenomenon known as *structural priming*, *syntactic persistence*, or *syntactic priming*. This effect has been widely studied as a window into how prior linguistic experience shapes subsequent behavior and what this reveals about the representations and memory mechanisms underlying language processing. Competing accounts attribute priming to residual activation, implicit learning, or interactions between the two (Bock, 1986b,a; Chang et al., 2000; Bock and Griffin, 2000; Fine and Jaeger, 2013).

Structural priming has been documented across many constructions and languages. Much of this evidence is consistent with accounts in which priming reflects sensitivity to local structural configurations, such as the arrangement of immediate phrasal constituents. Classic examples include the dative alternation and the active–passive alternation (Bock, 1986b; Pickering and Branigan, 1998; Bock and Loebell, 1990), along with a range of other constructions (e.g., Cleland and Pickering, 2003; Ferreira, 2003; Griffin and Weinstein-Tull, 2003; Hartsuiker and Westenberg, 2000). At the same time, more recent work suggests that priming is not limited to such local configurations. Studies of relative clause attachment ambiguity show priming between interpretations that differ only in hierarchical attachment, not in lexical content or linear order, suggesting that priming can also reflect sen-

sitivity to abstract structural relations (Desmet and Declercq, 2006; Scheepers, 2003).

Aligned with this view, cross-domain studies suggest that structural priming may extend beyond language itself. Scheepers et al. (2011) and Pozniak et al. (2018) showed that solving arithmetic expressions with different hierarchical groupings can bias subsequent relative clause attachment preferences in sentence processing. Because equations and sentences share no lexical or semantic content, these findings are difficult to explain in terms of lexical overlap or surface similarity. Instead, they have been interpreted as evidence that priming operates over abstract representations of hierarchical organization maintained in working memory (Scheepers et al., 2011; Pozniak et al., 2018).

Taken together, research on local configurations, global attachment, and cross-domain alignment suggests that structural priming reflects sensitivity to structured representations that extend beyond immediate word-order patterns.

2.2 Structural sensitivity in language models

Language model research has investigated whether these systems are sensitive to abstract linguistic structure, rather than relying only on surface-level distributional patterns. Researchers have approached this in several ways. For example, some studies have examined whether a model's internal representations encode syntactic dependencies or constructional information (Hewitt and Manning, 2019; Li et al., 2023; Tayyar Madabushi et al., 2020; White et al., 2021), while others have used targeted syntactic evaluation to test whether the model distinguishes between minimally contrasting grammatical and ungrammatical sentences (Gauthier et al., 2020; Marvin and Linzen, 2018b; Hu et al., 2020). Findings from this work suggest that language models can learn grammatical patterns that go beyond simple word-to-word associations. In other words, their behavior often reflects more than mere memorization of surface sequences. However, this evidence is still not fully conclusive. Showing that syntactic information can be recovered from a model's internal states does not necessarily mean that the model actively uses that information during prediction (Voita and Titov, 2020). Likewise, good performance on targeted syntactic evaluations does not always prove that the model is relying on abstract syntactic structure, since it may instead succeed by exploiting shallower lexical or distributional regularities. In addition, other work has pointed to weaknesses in areas such as word order, reliance on spurious heuristics, and the interpretation of negation (Kassner and Schütze, 2020; Lovering et al., 2021; Sinha et al., 2021). These findings suggest that language models do show some evidence of structural knowledge, but the depth, robustness, and functional role of that knowledge remain open questions.

2.3 Language models as developmental learners

The present study also connects to computational developmental linguistics, which treats learning systems as objects of developmental analysis rather than only as static predictors. From this perspective, language models are not simply evaluated for whether they know a particular syntactic contrast, but for what their behavior reveals about the emergence, organization, and limits of linguistic knowledge after large-scale training. This is especially relevant for cross-domain priming. If a model shows structural transfer from arithmetic to language, this would suggest that training has produced representations that abstract away from domain-specific surface forms. If it does not, the result suggests that linguistic and mathematical abilities may develop as partially or fully independent competencies, even in models that perform well in both domains. Thus, the present paradigm provides a developmental diagnostic: it asks whether exposure to multiple structured symbol systems leads to shared hierarchical representations or to functionally modular behavior across domains.

2.4 Structural priming in language models

Structural priming offers another way to study syntactic knowledge in language models. Prior work shows that language models do exhibit priming-like effects, including effects of recency, cumulative exposure, lexical overlap, and crosslinguistic persistence. At the same time, these effects do not always pattern like human priming: some studies report asymmetries across structural alternatives whose direction does not match the human literature. Structural priming is therefore informative not only because it provides further evidence that language models encode syntax, but also because it helps characterize how that knowledge is organized and how it may differ from human sentence processing.

Recent work including (Jumelet et al., 2024b; Sinclair et al., 2022b, 2026) has strengthened this

picture by showing that priming in neural language models cannot be reduced to simple surface repetition. In particular, (Sinclair et al., 2022b) reports reliable priming effects across a range of constructions and Transformer-based models, even when lexical overlap and semantic similarity are controlled. Priming is nevertheless sensitive to familiar psycholinguistic factors, weakening with distance and strengthening with repeated exposure and lexical or semantic similarity. This suggests that language models retain structural information across inputs, but that it remains closely tied to lexical, semantic, and distributional properties.

This raises a key question: ***Do priming effects in language models reflect abstract structural representations, or are they driven by domain-specific regularities?*** This question is especially pressing because existing work has focused almost entirely on within-language priming. It therefore remains unclear whether the representations supporting priming generalize across domains. The present study addresses this issue by asking whether exposure to structured input in mathematics influences the processing of structurally analogous input in language. Mathematics provides a particularly stringent test case because it is highly structured and compositional, yet differs sharply from language in surface form, semantics, and training distribution. If structural priming extends across these domains, that would support the existence of more domain-general structural representations in language models. If it does not, that would suggest important limits on the abstractness of their syntactic knowledge.

3 Experiments

3.1 Language Models

We evaluated five open-weight language models spanning two inference profiles: standard (non-reasoning) and reasoning (chain-of-thought), and two architectures, dense and mixture-of-experts (MoE). Table 1 summarizes the models and their key properties.

Models were selected to vary along two dimensions relevant to the study. First, we contrast **standard** models, which generate responses directly, with **reasoning** models, which produce an internal chain-of-thought before outputting the final answer. This distinction is important because structural priming would require the model to process the equation's hierarchical structure; reasoning models may engage with that structure more deeply due to their step-by-step computation. Second, we include both **dense** architectures (all parameters active on every token) and **mixture-of-experts** (MoE) architectures (only a subset of parameters active per token), allowing us to assess whether the total parameter count or the active computation determines math accuracy and attachment behavior.

All models were accessed via the Fireworks AI inference API.[2] Qwen3's default thinking mode was used and the <think> tags stripped before processing outputs, and GPT-OSS models were set to medium thinking effort. Experimental runs were logged with Langfuse for reproducibility.

3.2 Stimuli

Pozniak et al. (2018) and Scheepers et al. (2011) found that mathematical expressions can influence relative clause attachment in English and French. Building on this work, Experiment 1 adopted their stimulus materials in a Pozniak-style priming experiment. Specifically, we used 60 experimental item sets in Pozniak et al. (2018) (30 item sets in each language) to enable direct comparison with the human results. Each item consisted of two types of prime equations (2) and a written sentence as the target material (3).

(2) a. $90 - (9 + 1) \times 5$ HA prime

b. $90 - 9 + 1 \times 5$ LA prime

(3) Voici le tailleur de l'architecte qui
here the tailor of the architect who
s'apprête à créer un chef d'œuvre.
is about to create a masterpiece
'Here we have the tailor of the architect who is about to create a masterpiece.'

Note that we began with the English translations provided alongside the French stimuli and subsequently refined them to more closely match the French originals. Because the French materials encode grammatical gender on nouns, we took care to minimize any additional effects of gender and pronoun resolution in the English stimuli. Specifically, we replaced gender-marked forms with gender-neutral alternatives where possible, for example using *the* in place of *his* or *her* and *I* in place of *he* or *she*.

We then extended the investigation to a broader multilingual setting in Experiment 2, a MultiWho

[2] https://fireworks.ai

Model	Developer	Total	Active	Type
Llama 3.3 70B Instruct	Meta	70B	70B	Standard
Qwen3 8B	Alibaba	8B	8B	Reasoning
Kimi K2 Instruct	Moonshot AI	1T	32B	Standard (MoE)
gpt-oss-120b	OpenAI	117B	5.1B	Reasoning (MoE)
gpt-oss-20b	OpenAI	21B	3.6B	Reasoning (MoE)

Table 1: Models used in the experiments. *Total* = total parameter count; *Active* = parameters activated per forward pass (relevant for MoE models). Model cards are available on Hugging Face.[1]

multilingual extension, using the open-source *MultiWho* dataset (Lee et al., 2025), which contains ambiguous relative-clause attachment sentences from multiple languages. This extension allowed us to move beyond the smaller initial stimulus set based on the Pozniak-style materials and to examine whether the attachment patterns observed there would generalize to a larger and more cross-linguistically diverse set of items. The *MultiWho* dataset provides 96 ambiguous relative-clause sentences per language, substantially increasing the number of items relative to the initial experiment. To preserve comparability, we used the ambiguous *MultiWho* sentence materials together with arithmetic primes constructed in the same general format as those in the earlier experiment, thereby retaining the same priming logic and task structure while substantially expanding the language coverage and item base.

3.3 Procedure

In the baseline attachment condition, the model was presented only with ambiguous sentences in order to assess whether it exhibited an attachment preference and whether that preference aligned with previously reported human patterns. On each trial, the model received an ambiguous sentence such as (3), followed by a comprehension question (e.g., *Who is about to create a masterpiece?*).

In the priming condition, we tested whether arithmetic structure influenced subsequent attachment decisions. Each trial contained two components: (i) an arithmetic prime and (ii) an ambiguous sentence target followed by a comprehension question (Figure 3). We used an open-response format, allowing the model to generate its own answers to both tasks.

For Experiment 1, the priming materials consisted of 30 items, each repeated five times at a sampling temperature of 0.7. The design crossed two equation types (HA prime: *parenthesized*; LA prime: *flat*) with two target-language conditions

Combined condition prompt (open-response):

```
Answer both questions.  Give only short
answers, no explanation.
4 + (6 - 2) / 2 = ?

"Here we have the son of the father who
fancies riding what I like best."
Who fancies riding?

Respond in the format:
Math: [number]
Language: [single word]
```

Figure 3: Example prompt for the combined condition. The equation prime (here, a parenthesized/HA prime) and the ambiguous RC attachment sentence are presented together.

(English and French), yielding $30 \times 2 \times 2 \times 5 = 600$ combined trials. Together with the two baseline conditions, this resulted in 1200 trials per model: 300 math-baseline trials, 300 attachment-baseline trials, and 600 combined math+attachment trials.

Experiment 2 followed the same general procedure using the *MultiWho* dataset. It included 96 items in each of six languages, with five repetitions per item at a sampling temperature of 0.7. This yielded 2,880 attachment-baseline trials per model $(96 \times 6 \times 5)$ and 5,760 combined trials per model $(96 \times 6 \times 2 \times 5)$, for a total of 8,640 trials per model and 43,200 trials across the five models. All models were accessed through the Fireworks AI API with Langfuse tracing.

3.4 Analysis

Following standard practice in studies of relative clause attachment ambiguity, we use the proportion of HA responses as the primary measure. Because LA responses are simply the complement, values above 0.5 indicate an HA preference and values below 0.5 an LA preference.

Models did not always select one of the two candidate referents. Responses that matched neither the HA noun nor the LA noun were treated as invalid and excluded from HA/LA proportion calculations, though their counts are reported for transparency.

In the priming analysis, we further excluded trials with incorrect math answers. Following (Pozniak et al., 2018), only trials with correct math responses were retained, since incorrectly solved equations cannot be assumed to instantiate the intended structure.

3.5 Results in Exp. 1 (Pozniak-stimuli)

3.5.1 Baseline preference

As reported in the psycholinguistic literature, English generally shows a LA preference, whereas French shows a HA preference. The models only partially reproduced this cross-linguistic contrast (Table 2). GPT-OSS 120B and GPT-OSS 20B aligned most closely with the human pattern, showing lower HA rates in English and higher HA rates in French. Llama 3.3 70B Instruct showed the same directional contrast, but its English responses reflected only a weak LA preference. By contrast, Qwen3 8B and Kimi K2 Instruct favored HA in both languages.

The GPT-OSS results, however, require caution because these models produced unusually high rates of invalid responses, especially in English. This substantially reduced the number of valid English trials, leaving only 82 for GPT-OSS 120B and 43 for GPT-OSS 20B out of 150. Thus, although their overall directional pattern is consistent with the human literature, their English attachment estimates should be interpreted cautiously.

3.5.2 Priming Results

The priming results for English and French are summarized in Tables 3 and 4. We begin with English.

Table 3 shows no clear evidence of the predicted priming effect in English. Rather than shifting with prime type, the models largely maintained their baseline attachment preferences across conditions. Qwen 8B and Kimi K2 showed nearly identical HA rates across conditions, GPT-OSS 120B and GPT-OSS 20B remained broadly LA-preferring, and Llama 70B was difficult to interpret because very low math accuracy left too few valid trials.

The French results in Table 4 show the same pattern. Again, the models largely maintained their baseline preferences across prime conditions: Qwen 8B and Kimi K2 showed very similar HA rates across conditions, GPT-OSS 120B and GPT-OSS 20B remained broadly stable, and Llama 70B again yielded too few valid trials because of very low math accuracy.

Thus, neither the English nor the French results provide evidence of a priming effect. In both languages, attachment responses remained largely stable across prime conditions, suggesting that arithmetic structure did not systematically influence attachment choices in the predicted direction.

3.6 Results in Exp. 2 (multilingual extension)

We used the *MultiWho* dataset to extend the study to a larger multilingual set of relative-clause attachment items and to test whether the patterns from Experiment 1 generalize to this dataset.

3.6.1 Baseline Preference

We first examined baseline attachment preferences. Prior psycholinguistic research reports a LA preference in English and Chinese, and a HA preference in Japanese, Korean, Spanish, and Russian. The *MultiWho* baseline results show only partial alignment with these human patterns. English and Chinese were consistently LA-preferring across all models, in line with the human literature, and Russian showed a robust HA preference across models. Spanish showed weaker and less consistent evidence of HA. By contrast, Japanese and Korean did not show the expected HA preference: all models remained below 50% HA in both languages, although Qwen 8B, GPT-OSS 120B, and GPT-OSS 20B showed somewhat higher HA rates, especially in Japanese.

The clearest correspondence to the human literature was found for English, Chinese, and Russian, while Japanese and Korean diverged most clearly from the expected HA pattern. Spanish occupied an intermediate position, showing some evidence of HA but not a consistent human-like pattern across models.

Invalid responses also varied substantially across languages and models. Japanese showed the highest invalid-response rates, with especially high rates for GPT-OSS 20B and Kimi K2, and Korean also yielded elevated invalid-response rates for several models. By contrast, invalid responses were generally low in English, Spanish, and Chinese, aside from the GPT-OSS models in English.

3.6.2 Priming Results

The priming results for Experiment 2 are summarized in Table 6. If the arithmetic primes influenced attachment decisions, models should have produced more HA responses following HA primes and more LA responses following LA primes. This

		Llama 70B	Qwen 8B	Kimi K2	GPT-OSS 120B	GPT-OSS 20B
Preference	English	48.7% (73/150)	58.4% (80/137)	60.3% (88/146)	20.7% (17/82)	32.6% (14/43)
	French	73.3% (85/116)	92.2% (130/141)	67.8% (97/143)	67.8% (97/143)	84.3% (113/134)

Table 2: Baseline attachment preferences in Pozniak stimuli. Each model completed 300 trials. Percentages indicate HA responses among valid responses; fractions show HA/valid. Invalid counts are reported in Appendix B.

Condition	Llama 70B	Qwen 8B	Kimi K2	GPT-OSS 120B	GPT-OSS 20B
Math accuracy (HA prime)	18.0% (27/150)	100% (150/150)	99.3% (149/150)	100% (150/150)	100% (145/145)
Math accuracy (LA prime)	13.3% (20/150)	100% (150/150)	82.7% (124/150)	100% (150/150)	99.3% (140/141)
HA rate after HA prime	37.0% (10/27)	63.4% (92/145)	76.1% (105/138)	33.6% (41/122)	52.5% (52/99)
HA rate after LA prime	15.0% (3/20)	64.1% (91/142)	76.4% (84/110)	29.4% (32/109)	38.0% (30/79)
Priming effect (Δ)	+22.0	−0.7	−0.3	+4.2	+14.5

Table 3: English priming results (Pozniak stimuli). HA rate = proportion of high-attachment responses among valid trials after excluding incorrect math trials. Δ = HA rate after HA prime minus after LA prime (percentage points).

pattern was not observed consistently across languages or models. Instead, as in Experiment 1, the responses largely reflected the models' baseline attachment preferences in each language.

Regardless of prime type, in English, Chinese, Japanese, and Korean, where baseline preferences were generally low attachment, responses were dominated by LA choices. In Russian, where baseline preferences were strongly high attachment, responses were dominated by HA choices. Spanish showed a more mixed pattern across models, but again did not provide clear evidence that prime type systematically shifted attachment in the predicted direction. Thus, the *MultiWho* results provide no clear evidence of structural priming.

4 General Discussion

The present study investigated whether structural priming can be observed across domains in large language models, specifically from arithmetic expressions to relative-clause attachment. From a developmental perspective, this provides a test of whether linguistic and mathematical abilities in models rely on shared structural representations or remain functionally separate.

Across both experiments, we found no robust evidence of the predicted priming effect. If language models represented structure in a sufficiently abstract and transferable way across domains, HA primes should have increased HA responses and LA primes should have increased LA responses. This pattern did not emerge consistently. Instead, the models' responses largely reflected their baseline attachment preferences in each language. The absence of such transfer suggests that, at least under the present task conditions, model development does not necessarily yield domain-general hierarchical representations comparable to those implicated in human cross-domain priming. Rather, the models appear to develop structure-sensitive behavior that is more strongly tied to the domain, format, and task in which that structure is encountered.

The absence of a priming effect in both Experiment 1 and Experiment 2, including the larger multilingual *MultiWho* dataset, suggests that this null result is robust rather than simply an artifact of the smaller initial study. At the same time, the models were not entirely insensitive to structure: across both experiments, they showed clear attachment preferences that varied across languages. The crucial finding, however, is that these preferences were not systematically shifted by the arithmetic primes. This suggests that, although the models encode some structural information, that knowledge may not support abstract transfer across domains. More cautiously, their behavior may instead reflect domain-specific statistical regularities, learned associations, or task-specific processing biases rather than a shared abstract structural representation.

One possible interpretation of the null priming effect concerns the modularity, or functional separation, of linguistic and mathematical reasoning in language models. The models tested here were able, in many cases, to solve the arithmetic problems and to answer the attachment questions, indicating that failure to observe priming cannot be reduced simply to a complete inability to perform either task. However, successful performance in both domains does not entail that the same representations or processing routines are used across them.

Condition	Llama 70B	Qwen 8B	Kimi K2	GPT-OSS 120B	GPT-OSS 20B
Math accuracy (HA prime)	19.3% (29/150)	100% (150/150)	99.3% (149/150)	100% (150/150)	100% (150/150)
Math accuracy (LA prime)	11.3% (17/150)	100% (150/150)	85.3% (128/150)	100% (150/150)	100% (150/150)
HA rate after HA prime	69.0% (20/29)	84.4% (108/128)	79.9% (119/149)	70.0% (105/150)	87.3% (124/142)
HA rate after LA prime	29.4% (5/17)	87.5% (112/128)	81.9% (104/127)	65.5% (99/150)	86.8% (125/144)
Priming effect (Δ)	+39.6	−3.1	−2.0	+4.5	+0.5

Table 4: French priming results (Pozniak stimuli). HA rate = proportion of high-attachment responses among valid trials after excluding incorrect math trials. Δ = HA rate after HA prime minus after LA prime (percentage points).

Language	Llama 70B	Qwen 8B	Kimi K2	GPT-OSS 120B	GPT-OSS 20B
EN	13.2% (62/471)	18.6% (89/479)	27.1% (129/476)	7.6% (34/446)	12.6% (53/421)
CH	7.5% (35/469)	22.7% (106/467)	12.9% (61/474)	19.8% (94/474)	12.5% (57/456)
JP	13.0% (56/430)	38.2% (152/398)	23.5% (88/375)	35.6% (143/402)	36.3% (130/358)
KO	8.3% (34/411)	29.9% (134/448)	13.5% (56/414)	32.8% (151/461)	27.9% (124/445)
RU	57.4% (264/460)	73.8% (335/454)	63.4% (287/453)	77.4% (325/420)	76.9% (320/416)
SP	42.8% (196/458)	55.7% (264/474)	49.4% (235/476)	37.9% (180/475)	63.7% (297/466)

Table 5: Baseline attachment preferences in Experiment 2 (*MultiWho*). Each model completed 480 trials per language. Percentages indicate HA responses among valid responses, with raw counts in parentheses.

The absence of cross-domain priming is therefore compatible with the possibility that linguistic attachment preferences and arithmetic grouping are handled by partially separate mechanisms, representations, or task-specific circuits within the model. On this interpretation, mathematical and linguistic competence may coexist in the same model without being integrated at the level of abstract hierarchical structure required for priming.

This possibility is theoretically important rather than merely methodological. In humans, cross-domain priming has been interpreted as evidence for domain-general resources involved in representing hierarchical structure. The present findings suggest that language models may differ from humans not only in the amount or type of input they receive, but also in how competencies acquired from different input domains are organized. Thus, the null result contributes to a developmental account of model cognition: exposure to multiple structured domains may be sufficient for task performance, but not sufficient for the emergence of shared, transferable structural representations.

A second important finding concerns differences across models. Although the overall null priming result was similar across models, the models differed substantially in how reliably they carried out the task. Llama 3.3 70B Instruct consistently showed much lower math accuracy than the other models in the combined prime-and-sentence task, leaving relatively few interpretable trials for the priming analysis. The GPT-OSS models also showed distinctive response patterns, including relatively high invalid-response rates in some conditions, especially in English in Experiment 1. By contrast, Qwen3 8B and Kimi K2 Instruct were generally more stable in task execution, yielding a larger number of interpretable trials for analysis. These differences are important because they show that task reliability varied across models. However, this variation does not alter the broader conclusion: regardless of differences in execution, none of the models showed robust evidence of structural priming across domains.

A third point concerns baseline attachment preferences. Across both experiments, the models showed only partial alignment with human cross-linguistic patterns. In Experiment 1, some models captured the English–French contrast more clearly than others. In Experiment 2, the pattern was again mixed: English and Chinese generally showed LA preferences, and Russian showed a clear HA preference, but Japanese and Korean did not show the expected HA preference. This pattern is broadly consistent with prior *MultiWho* findings (Lee et al., 2025), suggesting that even newer models remain only partially sensitive to human-like cross-linguistic attachment preferences. At the same time, the English baseline preferences differed across Experiment 1 and Experiment 2, which likely reflects differences in the materials themselves. Compared with the *MultiWho* stimuli in Experiment 2, the relative-clause materials in Experiment 1 were longer and structurally heavier.

Language	Condition	Llama 70B	Qwen 8B	Kimi K2	GPT-OSS 120B	GPT-OSS 20B
EN	HA rate after HA prime	11.3%	25.3%	30.2%	6.8%	18.1%
	HA rate after LA prime	7.5%	24.8%	29.6%	4.2%	14.6%
	Δ	+3.8	+0.5	+0.6	+2.6	+3.5
CH	HA rate after HA prime	12.3%	22.2%	11.0%	16.6%	15.0%
	HA rate after LA prime	6.4%	19.4%	9.5%	17.6%	14.0%
	Δ	+5.9	+2.8	+1.5	−1.0	+1.0
JP	HA rate after HA prime	8.7%	29.7%	26.4%	34.8%	32.9%
	HA rate after LA prime	16.2%	29.3%	26.1%	37.1%	33.1%
	Δ	−7.5	+0.4	+0.3	−2.3	−0.2
KO	HA rate after HA prime	14.3%	19.6%	19.1%	30.6%	20.5%
	HA rate after LA prime	15.9%	20.2%	14.1%	31.5%	19.1%
	Δ	−1.6	−0.6	+5.0	−0.9	+1.4
RU	HA rate after HA prime	55.7%	60.2%	55.9%	72.6%	74.2%
	HA rate after LA prime	61.8%	61.5%	55.4%	73.4%	71.6%
	Δ	−6.1	−1.3	+0.5	−0.8	+2.6
SP	HA rate after HA prime	28.0%	45.8%	46.1%	31.6%	56.8%
	HA rate after LA prime	43.5%	47.0%	43.0%	31.6%	59.5%
	Δ	−15.5	−1.2	+3.1	0.0	−2.7

Table 6: Summary of priming results in the *MultiWho* dataset. HA rate = proportion of high-attachment responses among valid, math-correct responses. Δ = HA rate after HA prime − HA rate after LA prime (in percentage points). A positive Δ would indicate priming in the predicted direction. Across 30 language–model combinations, no consistent priming pattern emerges. (See Appendix C for full per-language tables with raw counts.)

Prior psycholinguistic work has shown that attachment preferences are influenced by properties of the input, including constituent length, processing complexity, and the distribution of lexical and structural cues (Hemforth et al., 1996). Similar considerations are relevant for language models, whose responses are often sensitive to surface form, input length, and local distributional patterns. From this perspective, the cross-experiment difference in English baseline preferences is not simply noise, but further evidence that attachment behavior in language models is shaped by the specific properties of the stimulus materials.

Taken together, these findings support a cautious but clear conclusion: although the models showed stable attachment preferences and some language-specific variation, they provided no robust evidence of structural priming from arithmetic expressions to relative-clause attachment. This suggests limited cross-domain structural generalization, with model behavior shaped more by baseline attachment tendencies and the statistical and formal properties of the linguistic input.

More broadly, the present results contribute to an ongoing debate about the nature of structure in language models. Prior work has shown that models can display sensitivity to syntactic patterns and can sometimes reproduce human-like preferences in linguistic tasks. The present findings qualify that picture by showing that such sensitivity does not necessarily extend to abstract structural priming across domains. In this respect, the study highlights an important distinction between exhibiting structured behavior within a domain and deploying abstract structural representations flexibly across domains. The models appear capable of the former, but the present evidence offers little support for the latter.

5 Conclusion

Across two experiments, we found no robust evidence of cross-domain structural priming from arithmetic expressions to relative-clause attachment in large language models. This suggests that, although these models show some structure-sensitive behavior within language, they do not readily generalize abstract structural representations across domains. The results point to a distinction between acquiring competence in multiple structured domains and using that competence in a way that supports cross-domain structural transfer. The absence of priming is therefore consistent with the possibility that linguistic and mathematical reasoning remain functionally separate in current models, though further representational analyses would be needed to test this interpretation directly.

6 Limitations

Several limitations of the present study should be noted. First, in some cases the combined math-and-language task yielded only a small number of usable trials because models either answered the math problem incorrectly or produced invalid responses to the attachment question. This was especially the case for Llama 3.3 70B Instruct, and for some language conditions with elevated invalid-response rates. Although these cases do not change the overall pattern of results, the small number of usable trials makes those estimates less stable for the specific model. As a result, these model-specific patterns should be interpreted with caution and should not be overgeneralized.

Second, the baseline attachment preferences differed across the two experiments, particularly in English. As discussed above, this difference may reflect differences in the stimulus materials, including sentence length, structural complexity, and other lexical or distributional properties. As a result, comparisons across datasets should be interpreted carefully, since attachment behavior may be shaped not only by language but also by properties of the items themselves.

Third, the present design tested a particularly strong form of structural generalization across domains, from arithmetic expressions to relative-clause attachment. Because the prime and target belong to different domains and differ substantially in surface form, this design places a demanding burden on the models. A null result in this setting therefore should not be taken to rule out the possibility that language models might show priming more readily in within-domain designs or in tasks with a more direct structural correspondence between prime and target.

A related limitation is that the present design cannot determine the internal source of the observed domain separation. The absence of priming may reflect genuinely modular or partially modular organization between linguistic and mathematical reasoning in the models. Alternatively, it may reflect properties of the prompting format, the open-response task, the salience of the arithmetic structure, or the way instruction-tuned models allocate attention across multi-part prompts. Thus, the present results should not be interpreted as proving architectural modularity in a strong sense. Rather, they show a functional absence of cross-domain transfer in this paradigm. Future work could test this interpretation more directly by examining intermediate representations, attention patterns, layer-wise effects, or developmental checkpoints during training to determine whether linguistic and mathematical structure become more integrated over time.

References

J Kathryn Bock. 1986a. Meaning, sound, and syntax: Lexical priming in sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12(4):575.

J Kathryn Bock. 1986b. Syntactic persistence in language production. *Cognitive psychology*, 18(3):355–387.

Kathryn Bock and Zenzi M Griffin. 2000. The persistence of structural priming: Transient activation or implicit learning? *Journal of experimental psychology: General*, 129(2):177.

Kathryn Bock and Helga Loebell. 1990. Framing sentences. *Cognition*, 35(1):1–39.

Zhenguang Cai, Xufeng Duan, David Haslett, Shuqi Wang, and Martin Pickering. 2024. Do large language models resemble humans in language use? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 37–56, Bangkok, Thailand. Association for Computational Linguistics.

Franklin Chang, Gary S Dell, Kathryn Bock, and Zenzi M Griffin. 2000. Structural priming as implicit learning: A comparison of models of sentence production. *Journal of psycholinguistic research*, 29(2):217–230.

Alexandra A Cleland and Martin J Pickering. 2003. The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language*, 49(2):214–230.

Timothy Desmet and Mieke Declercq. 2006. Cross-linguistic priming of syntactic hierarchical configuration information. *Journal of Memory and Language*, 54(4):610–632.

Victor S Ferreira. 2003. The persistence of optional complementizer production: Why saying "that" is not saying "that" at all. *Journal of Memory and Language*, 48(2):379–398.

Alex Fine and T Florian Jaeger. 2013. Syntactic priming in language comprehension allows linguistic expectations to converge on the statistics of the input. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35.

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, Volume 1 (Long and Short Papers)*, 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.

Zenzi M Griffin and Justin Weinstein-Tull. 2003. Conceptual structure modulates structural priming in the production of complex sentences. *Journal of Memory and Language*, 49(4):537–555.

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.

Robert J Hartsuiker and Casper Westenberg. 2000. Word order priming in written and spoken sentence production. *Cognition*, 75(2):B27–B39.

B Hemforth, L Konieczny, and C Scheepers. 1996. Syntactic and anaphoric processes in modifier attachment. In *The 9th Annual CUNY Conference on Human Sentence Processing*, pages 21–23.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. 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.

Jaap Jumelet, Willem Zuidema, and Arabella Sinclair. 2024a. Do language models exhibit human-like structural priming effects? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14727–14742, Bangkok, Thailand. Association for Computational Linguistics.

Jaap Jumelet, Willem Zuidema, and Arabella Sinclair. 2024b. Do language models exhibit human-like structural priming effects? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14727–14742.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7811–7818.

So Young Lee, Russell Scheinberg, Amber Shore, and Ameeta Agrawal. 2025. Who relies more on world knowledge and bias for syntactic ambiguity resolution: Humans or LLMs? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3484–3498, Albuquerque, New Mexico. Association for Computational Linguistics.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023. Generative judge for evaluating alignment. *Preprint*, arXiv:2310.05470.

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.

Charles Lovering, Rohan Jha, Tal Linzen, and Ellie Pavlick. 2021. Predicting inductive biases of pretrained models. In *International Conference on learning representations*.

Rebecca Marvin and Tal Linzen. 2018a. 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.

Rebecca Marvin and Tal Linzen. 2018b. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 1192–1202.

James A. Michaelov, Catherine Arnett, Tyler A. Chang, and Benjamin K. Bergen. 2023. Structural priming demonstrates abstract grammatical representations in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3703–3720, Singapore. Association for Computational Linguistics.

Martin J Pickering and Holly P Branigan. 1998. The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and language*, 39(4):633–651.

Céline Pozniak, Barbara Hemforth, and Christoph Scheepers. 2018. Cross-domain priming from mathematics to relative-clause attachment: A visual-world study in french. *Frontiers in psychology*, 9:2056.

Grusha Prasad, Marten van Schijndel, and Tal Linzen. 2019. Using priming to uncover the organization of syntactic representations in neural language models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages

66–76, Hong Kong, China. Association for Computational Linguistics.

Christoph Scheepers. 2003. Syntactic priming of relative clause attachments: Persistence of structural configuration in sentence production. *Cognition*, 89(3):179–205.

Christoph Scheepers, Patrick Sturt, Catherine J Martin, Andriy Myachykov, Kay Teevan, and Izabela Viskupova. 2011. Structural priming across cognitive domains: From simple arithmetic to relative-clause attachment. *Psychological Science*, 22(10):1319–1326.

Arabella Sinclair, Jaap Jumelet, Willem Zuidema, and Raquel Fernández. 2022a. Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

Arabella Sinclair, Jaap Jumelet, Willem Zuidema, and Raquel Fernández. 2022b. Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

Arabella Sinclair, Anastasia Klimovich-Gray, Jaap Jumelet, Nika Adamian, and Agnieszka Konopka. 2026. Structural priming in humans and large language models. *Journal of Memory and Language*, 149:104713.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 2888–2913.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets construction grammar. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4020–4032, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. 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.

Jennifer C White, Tiago Pimentel, Naomi Saphra, and Ryan Cotterell. 2021. A non-linear structural probe. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 132–138.

Ethan Wilcox, Clara Meister, Ryan Cotterell, and Tiago Pimentel. 2023. Language model quality correlates with psychometric predictive power in multiple languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7503–7511, Singapore. Association for Computational Linguistics.

A Appendix: English Stimulus Modifications

The English stimuli were adapted from the translations provided alongside the French originals in Pozniak et al. (2018). Because the French materials use possessives that agree with the possessed noun (e.g., *son vêtement* 'his/her garment'), they do not introduce gender as an unintended disambiguation cue. The English translations, however, contained gendered pronouns (*he*, *she*, *his*, *her*) that could bias attachment when NP1 and NP2 differ in gender. For example, in *"the daughter of the chemist who will put on* ***her*** *usual outfit"*, the pronoun *her* might more easily be construed as referring to the daughter.

To remove this confound, we replaced gendered pronouns with first-person forms (*he/she* → *I*, *his/her* → *my*) or rephrased to avoid pronouns entirely (*his drink* → *a drink*). Three items (6, 12, and 15) required no changes. Table 7 lists all modifications.

Item	Sentence
1	*...who fancies riding what ~~he likes~~* ***I like*** *best.*
2	*...who will set up what ~~he was~~* ***I was*** *asked to.*
3	*...who is about to have ~~his drink~~* ***a drink****.*
4	*...who will prepare what ~~he needs to~~* ***is needed****.*
5	*...who will cut what ~~he has to~~* ***is needed****.*
7	*...who will reveal ~~his~~* ***my*** *latest creation.*
8	*...who will read what ~~she's~~* ***I'm*** *used to.*
9	*...who will continue working on ~~her~~* ***my*** *latest project.*
10	*...who will deliver what ~~he has~~* ***I have****.*
11	*...who will present ~~his~~* ***my*** *recent work.*
13	*...who will sip ~~her favorite~~* ***a favorite*** *drink.*
14	*...who will fetch what ~~she's~~* ***I'm*** *used to.*
16	*...who will purchase what ~~he needs~~* ***I need****.*
17	*...who will create what ~~he's~~* ***I'm*** *best known for.*
18	*...who will fetch ~~her~~* ***my*** *favorite item.*
19	*...who will work on what ~~he's~~* ***I'm*** *supposed to.*
20	*...who will wear ~~her~~* ***my*** *favorite clothes.*
21	*...who will grab what ~~she~~* ***I*** *was looking for.*
22	*...who'll have to sell ~~his~~* ***my*** *favorite possession.*
23	*...who will eat ~~his lunch~~* ***lunch****.*
24	*...who will have ~~his dinner~~* ***dinner****.*
25	*...who will put on ~~his~~* ***my*** *usual outfit.*
26	*...who will finish what ~~he was~~* ***I was*** *working on.*
27	*...who will finish what ~~he started~~* ***I started*** *to work on.*
28	*...who will put on ~~her~~* ***my*** *usual outfit in the morning.*
29	*...who will read what ~~he's~~* ***I'm*** *most interested in.*
30	*...who will have what ~~he's~~* ***I'm*** *desperate for.*

Table 7: Modifications to the English stimuli from Pozniak et al. (2018). Only the relative clause portion is shown; the matrix clause (e.g., *"Here we have the son of the father..."*) was unchanged. Items 6, 12, and 15 required no modification.

B Appendix: Invalid Responses Rate

	Llama 70B	Qwen 8B	Kimi K2	GPT-OSS 120B	GPT-OSS 20B
English	0% (0/150)	9% (13/150)	3% (4/150)	45% (68/150)	71% (107/150)
French	23% (34/150)	6% (9/150)	5% (7/150)	5% (7/150)	11% (16/150)

Table 8: Invalid response counts by language in Experiment 1 (*Pozniak*). Each cell is based on 150 trials.

	Llama 70B	Qwen 8B	Kimi K2	GPT-OSS 120B	GPT-OSS 20B
English	2% (9/480)	0% (1/480)	1% (4/480)	7% (34/480)	12% (59/480)
Chinese	2% (11/480)	3% (13/480)	1% (6/480)	1% (6/480)	5% (24/480)
Japanese	10% (50/480)	17% (82/480)	22% (105/480)	16% (78/480)	25% (122/480)
Korean	14% (69/480)	7% (32/480)	14% (66/480)	4% (19/480)	7% (35/480)
Russian	4% (20/480)	5% (26/480)	6% (27/480)	13% (60/480)	13% (64/480)
Spanish	5% (22/480)	1% (6/480)	1% (4/480)	1% (5/480)	3% (14/480)

Table 9: Invalid response rates by language in Experiment 2 (*MultiWho*). Each cell shows the percentage of invalid responses, followed by the raw count in parentheses.

C MultiWho Priming Results Detail

Condition	Llama 70B	Qwen 8B	Kimi K2	GPT-OSS 120B	GPT-OSS 20B
Math accuracy (HA prime)	28.0% (134/480)	100% (480/480)	90.6% (435/480)	100% (480/480)	100% (474/474)
Sentence HA answer rate	11.3% (15/133)	25.3% (118/467)	30.2% (130/430)	6.8% (31/457)	18.1% (81/447)
Math accuracy (LA prime)	16.7% (80/480)	100% (480/480)	73.8% (354/480)	100% (480/480)	98.9% (466/471)
Sentence LA answer rate	92.5% (74/80)	75.2% (357/475)	70.4% (243/345)	95.8% (436/455)	85.4% (369/432)

Table 10: English priming results by model in the *MultiWho* dataset. Fractions in parentheses indicate raw counts. Percentages for HA and LA answers are calculated within valid attachment responses after excluding trials with incorrect math answers.

Condition	Llama 70B	Qwen 8B	Kimi K2	GPT-OSS 120B	GPT-OSS 20B
Math accuracy (HA prime)	25.4% (122/480)	100% (480/480)	91.0% (437/480)	100% (480/480)	100% (480/480)
Sentence HA answer rate	12.3% (15/122)	22.2% (102/460)	11.0% (46/419)	16.6% (78/469)	15.0% (66/441)
Math accuracy (LA prime)	18.1% (87/480)	100% (480/480)	75.4% (362/480)	100% (480/480)	99.0% (475/480)
Sentence LA answer rate	93.6% (73/78)	80.6% (378/469)	90.5% (306/338)	82.4% (383/465)	86.0% (368/428)

Table 11: Chinese priming results by model in the *MultiWho* dataset. Fractions in parentheses indicate raw counts. Percentages for HA and LA answers are calculated within valid attachment responses after excluding trials with incorrect math answers.

Condition	Llama 70B	Qwen 8B	Kimi K2	GPT-OSS 120B	GPT-OSS 20B
Math accuracy (HA prime)	27.1% (130/480)	100% (480/480)	90.8% (436/480)	100% (480/480)	100% (480/480)
Sentence HA answer rate	8.7% (9/103)	29.7% (114/384)	26.4% (66/250)	34.8% (149/428)	32.9% (105/319)
Math accuracy (LA prime)	19.4% (93/480)	100% (480/480)	76.9% (369/480)	100% (480/480)	99.4% (477/480)
Sentence LA answer rate	83.8% (57/68)	70.7% (265/375)	73.9% (153/207)	62.9% (264/420)	66.9% (216/323)

Table 12: Japanese priming results by model in the *MultiWho* dataset. Fractions in parentheses indicate raw counts. Percentages for HA and LA answers are calculated within valid attachment responses after excluding trials with incorrect math answers.

Condition	Llama 70B	Qwen 8B	Kimi K2	GPT-OSS 120B	GPT-OSS 20B
Math accuracy (HA prime)	27.9% (134/480)	100% (480/480)	91.9% (441/480)	100% (480/480)	100% (480/480)
Sentence HA answer rate	14.3% (16/112)	19.6% (83/424)	19.1% (61/319)	30.6% (141/461)	20.5% (79/386)
Math accuracy (LA prime)	19.4% (93/480)	100% (480/480)	73.3% (352/480)	100% (480/480)	99.8% (479/480)
Sentence LA answer rate	84.1% (69/82)	79.8% (344/431)	85.9% (214/249)	68.5% (315/460)	80.9% (321/397)

Table 13: Korean priming results by model in the *MultiWho* dataset. Fractions in parentheses indicate raw counts. Percentages for HA and LA answers are calculated within valid attachment responses after excluding trials with incorrect math answers.

Condition	Llama 70B	Qwen 8B	Kimi K2	GPT-OSS 120B	GPT-OSS 20B
Math accuracy (HA prime)	29.8% (143/480)	100% (480/480)	92.7% (445/480)	100% (480/480)	100% (480/480)
Sentence HA answer rate	55.7% (73/131)	60.2% (277/460)	55.9% (224/401)	72.6% (339/467)	74.2% (333/449)
Math accuracy (LA prime)	17.7% (85/480)	100% (480/480)	74.2% (356/480)	100% (480/480)	98.8% (474/477)
Sentence LA answer rate	38.2% (29/76)	38.5% (180/468)	44.6% (144/323)	26.6% (122/459)	28.4% (128/450)

Table 14: Russian priming results by model in the *MultiWho* dataset. Fractions in parentheses indicate raw counts. Percentages for HA and LA answers are calculated within valid attachment responses after excluding trials with incorrect math answers.

Condition	Llama 70B	Qwen 8B	Kimi K2	GPT-OSS 120B	GPT-OSS 20B
Math accuracy (HA prime)	29.8% (143/480)	100% (480/480)	91.7% (440/480)	100% (480/480)	100% (480/480)
Sentence HA answer rate	28.0% (40/143)	45.8% (216/472)	46.1% (195/423)	31.6% (151/478)	56.8% (266/468)
Math accuracy (LA prime)	17.7% (85/480)	100% (480/480)	73.8% (354/480)	100% (480/480)	99.2% (476/479)
Sentence LA answer rate	56.5% (48/85)	53.0% (249/470)	57.0% (195/342)	68.4% (324/474)	40.5% (189/467)

Table 15: Spanish priming results by model in the *MultiWho* dataset. Fractions in parentheses indicate raw counts. Percentages for HA and LA answers are calculated within valid attachment responses after excluding trials with incorrect math answers.

Do large language models and humans follow similar L2 learning stages? Assessing GPT-2's Swedish grammar acquisition within the Processability Theory framework

Stella Lundqvist[1] **Murathan Kurfalı**[2] **Johan Sjons**[1]
[1]Department of Linguistics and Philology, Uppsala University, Sweden
[2]RISE Research Institutes of Sweden
stellazoelouise.lundqvist.0915@student.uu.se murathan.kurfali@ri.se
johan.sjons@lingfil.uu.se

Abstract

We investigate whether GPT-2 acquires Swedish grammatical structures in the same implicational order as for human second language (L2) learners, as predicted by Processability Theory (PT). We present SwePT – a minimal pair dataset targeting Swedish syntactic and morphological structures that are acquired by human L2 learners on four separate stages of language development – and evaluate the GPT-2 models on SwePT using an acceptability classification task throughout fine-tuning with different input orders in regards to the grammatical structures identified in the data. We find that the observed acquisition orders correlate across the fine-tuned models, while violating the implicational order sequence as hypothesized by PT. The observed relation between performance on the classification task and frequency distributions of the contrasting features in the minimal pairs suggests that the acquisition order can be explained by unigram and n-gram heuristics. While the adaptation of NLP methodologies into the PT framework requires further conceptual and methodological refinement, we do not find evidence for PT-like grammatical development in our experiments.

1 Introduction

Despite language models' (LLMs) ability to acquire complex linguistic patterns and generate coherent and human-like language, the mechanisms underlying their grammatical development remain poorly understood. Research on these capabilities often draws inspiration from studies of child language development (e.g., McCoy et al., 2018; Choshen et al., 2021; Warstadt et al., 2023; Evanson et al., 2023; Yedetore et al., 2023), yet insights from second language acquisition (SLA) research remain largely unexplored. One promising framework is Processability Theory (PT; Pienemann, 1998b, 2005), one of the most influential theories on second language (L2) grammatical development in SLA research. This psycholinguistic theory posits that a learner can acquire only those linguistic forms and functions that they are cognitively prepared to process, constrained by the processing procedures available at their current stage of language development. The procedures are accessed in a predictable order sequence, supported by cross-linguistic evidence from numerous empirical studies of speech production and grammatical perception (e.g., Norrby and Håkansson, 2007; Kawaguchi, 2008; Mansouri, 2008; Ellis, 2008; Keatinge and Keßler, 2009; Wang, 2011; Spinner, 2013; Buyl and Housen, 2015). To date, however, no study has investigated the hypotheses of PT on artificial learners.

In this study, we evaluate an LLM against PT's developmental sequence, allowing us to examine whether the emergence of grammatical patterns in artificial systems follows trajectories similar to those observed in human second language learners. We approximate Swedish L2 learners by fine-tuning a GPT-2 model pretrained on English, on Swedish text data organized in three different curricula: randomized input order, order of increasing complexity and order of decreasing complexity, as hypothesized by PT.

We present the Swedish Processability Theory Minimal Pair Dataset (SwePT), consisting of nine subsets of minimal pairs containing different grammatical phenomena that represent four stages of the Swedish PT developmental hierarchy. We test the models' grammatical knowledge development using acceptability judgments (AJT) on SwePT at regular intervals throughout fine-tuning, to examine whether the model exhibits a developmental trajectory in acquiring Swedish grammatical structures that aligns with the trajectory that has been observed within human L2 learners. We evaluate the AJT using adapted implementations of the emergence criterion and implicational scaling, that are traditionally used to test human learner output within the PT framework, and analyze the learning

Proceedings of the 1st Workshop on Computational Developmental Linguistics (CDL), pages 52–76

trajectories.

The main contributions of this study are: (i) introducing a novel approach and bridging the fields of SLA and research on LLMs in using the PT framework to the evaluation of LLMs; (ii) creating and publicly releasing the *Swedish Processability Theory Minimal Pair Dataset* (SwePT) including canonical word order SVO, plural, tense, attributive agreement, predicative agreement (with and without attractor nouns), inversion after topicalization, preverbal negation and non-inversion in indirect questions – along with our codebase; and (iii) evaluating GPT-2 on its performance and acquisition order of the structures in SwePT using AJT with curriculum learning. By evaluating whether human developmental patterns are shared by artificial learners such as GPT-2, our study may contribute to the emerging research field of learning trajectories, to the expansion of the PT framework, and to our understanding of universal principles of language acquisition and how they differ between humans and artificial learners. Additionally, testing curriculum learning effects on grammatical development in artificial learners may contribute to the development of more efficient methods for training LLMs with input sequencing.

2 Background and Related Work

2.1 Learning trajectories

While learning trajectories in language models is a fairly new area of research, several recent studies have examined how and when language models acquire different linguistic phenomena during pretraining (Saphra and Lopez, 2018; Chiang et al., 2020; Liu et al., 2021; Choshen et al., 2021; Blevins et al., 2022; La Fiandra et al., 2025). Choshen et al. (2021) found a systematic learning trajectory across LLMs with different initializations, architecture and training data, albeit at different speeds, and that morphological phenomena was found to emerge at similar stages. In initial stages, the LLMs were found to rely on local cues such as the frequency of the preceding words, similarly to bag-of-words (BOW) models. Performance is high for tasks such as Part-of-speech (POS) tagging during this stage (Saphra and Lopez, 2018). This correlation subsides as training progresses, as the models seem to apply different strategies. For some simple linguistic structures, this change in strategies can cause accuracies that start high to drop (Choshen et al., 2021).

In later stages of training, the LLMs' accuracy scores correlate with those of n-gram models, suggesting that the models are relying less on simple frequencies and more on structural cues and global features. Simultaneously, syntactic depth becomes a greater predictor to performance than sentence length. As training progresses, the LLMs' performances become more similar to humans', eventually reaching a plateau (Choshen et al., 2021).

While the linguistic phenomena and their acquisition trajectories in these studies are not categorized in accordance with their hypothesized processability, the observed progression from local to global cues aligns with the progression across the developmental stages as described within the PT framework. This further motivates our study.

2.2 Measuring Linguistic Competence within the PT Framework

PT is built upon Levelt's (1989) model of speech *production*, which inherently views the cognitive processes of production and reception as separate. It is thus not surprising that the vast majority of PT studies concern speech production, with only a few (e.g. Norrby and Håkansson, 2007) including written data. Four studies (Ellis, 2008; Keatinge and Keßler, 2009; Spinner, 2013; Buyl and Housen, 2015) have previously investigated grammatical *comprehension* within the PT framework. The lack of unity in the methodological approaches and findings in these studies highlight the need for additional research to determine whether PT can predict receptive processing sequences with the same reliability as productive sequences.

2.3 AJT and Minimal Pair Benchmarks

A common method for inferring linguistic knowledge of language models is using acceptability judgment tests (AJT) with benchmarks of *minimal pairs*, where the learner is presented with one grammatical and one ungrammatical sentence that differ from each other on a single linguistic aspect and is tasked to determine which one of them is grammatical. Significant contributions to this practice include the Corpus of Linguistic Acceptability (CoLA, (Warstadt, 2019), The Benchmark of Linguistic Minimal Pairs for English (BLiMP Warstadt et al., 2020) and the Russian Benchmark of Linguistic Minimal Pairs (RuBLiMP, Taktasheva et al., 2024). The most significant contribution to Swedish AJT are the Dataset for Linguistic Acceptability Judgments (DaLAJ and DaLAJ-GED,

Volodina et al., 2021, 2023). While these benchmarks include a large number of minimal pairs and cover a broad range of linguistic phenomena allowing for evaluating general performance on AJT, SwePT is designed specifically to test the developmental sequence as predicted by PT, including linguistic phenomena not covered in pre-existing Swedish datasets.

3 Methodology

3.1 SwePT: A Swedish PT minimal pairs dataset

We present the Swedish Processability Theory Minimal Pair Dataset (SwePT), consisting of nine subsets of minimal pairs containing different grammatical phenomena that represent four stages of the Swedish PT developmental hierarchy, namely SVO (canonical word order SVO, 2nd stage), PLU (plural, 2nd stage), TNS (tense, 2nd stage), ATT (attributive agreement, 3rd stage), PR_a (predicative agreement, 4th stage), PR_b (predicative agreement with attractors, 4th stage), INV (inversion after topicalization, 4th stage), NEG (preverbal negation, 5th stage) and INQ (Non-inversion in indirect questions, 5th stage). Examples of the minimal pairs representing each subset are presented in Table 1.

Processing Pipeline. SwePT was constructed with an automated approach similar to that of RuBLiMP (Taktasheva et al., 2024). The grammatical sentences of each minimal pair were selected from the Swedish Talbanken and LinES treebanks from UD (De Marneffe et al., 2021) after processing the sentences through a custom pipeline identifying the target linguistic structures using rule-based Python scripts.[1] The scripts were written by performing several manual iterations of systematically relaxing the heuristics and reviewing the output. The criteria for identification and perturbation of the structures are found in Appendix A.

The pipeline performs three main consecutive steps: 1) identifying and extracting sentences containing the PT structures from the source CoNLL-U files through a dependency tree search, 2) duplicating the sentences to form the minimal pairs, and 3) altering the duplicates into ungrammatical sentences with respect to their target structures. The first step of this process was also used for labeling the training data (see Section 3.2). To form the minimal pairs of the syntactic structures (SVO, INV, INQ and NEG), relevant grammatical constituents and arguments were identified and had their positions switched with respect to the target structure. The alteration of the morphological structures (PLU, TNS, ATT and PR_a) was performed by converting the conjugated target structures into their neutral form (lemma). The alteration process for the PR_b minimal pairs was performed manually in order to minimize errors, due to the small amount of extracted sentences and the complexity of the alteration task. The details of the process are described in Appendix A.

3.2 Fine-tuning Data

For fine-tuning, we used the Swedish partition of the Common Crawl corpus Open Super-large Crawled Aggregated coRpus (OSCAR)[2]. Due to limited computational resources only 9% of the dataset (approx. 680k examples and 1B tokens) was extracted for the training set after shuffling the data (seed=42). The data was processed in multiple steps, with the objective to separate the data into four subsets representing stages 2-5 in the PT hierarchy. The parsing, labeling and grouping processes are described in the sections below.

Labeling the Fine-tuning Data. Each sentence in the data was first parsed, annotated and converted into CoNLL-U format using Stanza (Qi et al., 2020). The parsing performance was evaluated by comparing the distribution of linguistic categories in the parsed OSCAR subset with those in the gold-standard Talbanken and LinES corpora (Table 5 and Figure 2, Appendix B). After parsing, the CoNLL-U sentences were processed through the same functions used to identify the structures for the SwePT dataset, from which each sentence is returned labeled with the structures identified within it. 500k sentences were then randomly sampled from the labeled sentences in order to ensure that one epoch of training across the entire dataset would fit within 72 h of training (as calculated during a test run). See Appendix A for details on the identification criteria.

Grouping the Fine-tuning Data. After parsing the raw text from OSCAR and labeling the training data, the sentences were grouped into four subsets

[1]The scripts and datasets are available here: https://github.com/stellson/SwePT

[2]https://huggingface.co/datasets/oscar-corpus/OSCAR-2201

Structure	n of pairs	Example
5 NEG	303	Men det är viktigt, att förlusterna [inte] [blir] onödigt stora. *Men det är viktigt, att förlusterna [blir] [inte] onödigt stora. (*But it is important that the losses are not unnecessarily large.*)
5 INQ	94	Jag har lust att fråga honom varför [den] inte [trycktes]. *Jag har lust att fråga honom varför [trycktes] [den] inte. (*I want to ask him why it wasn't printed.*)
4 INV	2581	Ovanpå ett skåp i hörnet [satt] [Dobby] hopkrupen. *Ovanpå ett skåp i hörnet [Dobby] [satt] hopkrupen. (*On top of a cupboard in the corner crouched Dobby.*)
4 PR_a	226	De flesta u-länder har varit [koloniserade] *De flesta u-länder har varit [koloniserad] *(Most developing countries have been colonized)*
4 PR_b	27	Resultaten av uppväxten i denna miljö är rätt så [uppenbara]. *Resultaten av uppväxten i denna miljö är rätt så [uppenbar]. (*The results of growing up in this environment are quite obvious.*)
3 ATT	213	Han har inget [civiliserat] ansikte. *Han har inget [civiliserad] ansikte. (*He does not have a civilized face.*)
2 TNS	2000	Jag [är] min fars dotter. *Jag [vara] min fars dotter. (*I am my father's daughter.*)
2 PLU	479	Måste du försöka göra åtta [saker] samtidigt? *Måste du försöka göra åtta [sak] samtidigt? (*Must you try and do eight things at once?*)
2 SVO	2519	Hon [hade] [en dämpad, tonlös röst] och bröt inte så kraftigt som mannen. *Hon [en dämpad, tonlös röst] [hade] och bröt inte så kraftigt som mannen. (*She had a soft, dry voice and her accent was slighter than her husband's.*)

Table 1: Selected examples of minimal pairs (a grammatical sentence and its ungrammatical equivalent) from SwePT, including their translations. The target structures are displayed within square brackets.

representing each of the developmental stages (2–5) in the PT hierarchy. The subsets were populated in decreasing order, and the sentences in each subset thus only contain 1) structures from its respective stage, and 2) structures from lower stages, if occurring within the same sentences. The sentences that remained unlabeled after labeling (i.e., none of the target structures were identified within them) were distributed into the four subsets in proportion to the original size of each subset. The distribution between stages is shown in Table 4 in Appendix B. Observe that the subsets are different in size, since each developmental stage are represented by different numbers of linguistic structures that occur in varying frequencies in the training data.

3.3 Fine-tuning

Models. We fine-tuned and evaluated four small (124 M parameters) GPT-2 model instances[3] pre-trained on English. As a causal (unidirectional) transformer, GPT-2 estimates the probability of the next word given its previous context (Radford et al., 2019). This aspect is similar to the incremental processing of humans (e.g., Altmann and Kamide, 1999; Kuribayashi et al., 2025), which is one of the reasons that this model was chosen for this project. Another reason is the relatively small size, allowing for effective fine-tuning on a smaller dataset, which was crucial due to limited computing resources.

Curriculum learning. We employed the method of curriculum learning (Bengio et al., 2009) during fine-tuning, where models are initially fine-tuned

[3]https://huggingface.co/openai-community/gpt2

on simpler concepts and gradually move on to more complex concepts. We used three different curricula including one randomized input order, in order to test the robustness of the implicational acquisition order as stipulated by PT. We fine-tuned one model instance on input data ordered from simpler to more complex (`GPT-order`, seed=42), one in reverse order (`GPT-reverse`, seed=42) and two on all four subsets concatenated into one dataset (`GPT-mixed`, seed=42 and `GPT-mixed_2`, seed=123), thus exposing the models to a randomized curriculum. The models were trained for 72 hours for one epoch. If `GPT-reverse` displays a similar acquisition order as the other models, it is implicated that the implicational acquisition order as stipulated by PT holds. Checkpoints were saved at each 100th time step and named according to their indices. Training arguments are specified in Appendix B.3.

3.4 Evaluation

Acceptability Judgment Test. We follow the approach of Evanson et al. (2023) in conducting the AJT. At each checkpoint (every 100 training steps), we measure how acceptable the model finds each grammatical and ungrammatical sentence of each pair. More specifically, the score is calculated as follows, $-\mathcal{L}(M, X) \times N = \sum_{t=1}^{N} \log P(x_t \mid x_{<t}, M)$, where the total log-likelihood of a sentence S equals the cross-entropy loss $\mathcal{L}(M, X)$ (negative average log-likelihood) of N tokens in the sentence. The accuracy is calculated as the percentage of the pairs where the grammatical sentence was given a higher score than its ungrammatical counterpart.

Acquisition Time and the Emergence Criterion. While most SLA theories use native-like performance or accuracy as its metric for assessing grammatical knowledge, in PT studies the current level of the learner's language development is determined using the *emergence criterion* (Pienemann, 1998b). Emergence of a certain grammatical rule is represented by a learner's first production of a token of that rule, and marks the onset of the procedure that enables its acquisition. More specifically, the emergence criterion relies on consideration of four possible cases, namely (1) a lack of evidence (i.e. no present obligatory context for the target rule), (2) insufficient evidence (i.e. insufficient number of examples), (3) counter-evidence (i.e. non-application of the rule in the presence of its obligatory contexts) and (4) evidence of rule application (i.e., sufficient examples of applications of the rule in the presence of its obligatory context; see (Pienemann, 1998b).[4]

The emergence criterion has been adapted and reduced to case (3) (interpreted as higher average score assigned to the grammatical sentence) and (4) (interpreted as a higher average score assigned to the ungrammatical sentence), as AJT and not language output are used for measuring the models' acquisition in our study. Before reaching a certain threshold during training, the model is expected to distribute the probabilities over the grammatical and ungrammatical sentences somewhat randomly, and the model's acquisition of the structure cannot be inferred from a single correctly identified grammatical sentence without the context of the cumulative probabilities of the entire subset. To account for some noise around the chance level mark, we thus set the acquisition threshold at 60% accuracy. Following the approach of Buyl and Housen (2015), we also evaluated at 50% and 80%, and calculated the k number of sentences per subset that must be correct in order to ensure acquisition at each threshold. The acquisition thresholds are displayed in Table 6 in Appendix C.

Implicational Scaling. In order to test whether the acquisition of grammatical structures follows a hierarchical, implicational pattern across learners, as predicted by PT, implicational scaling (Rickford, 2004) is used. Implicational scales are binary matrices that visualize what structures are acquired by each learner at the time of evaluation. PT predicts only the order of acquisition and thus allows for variation in terms of the speed in which learners acquire the processing procedures as well as the order among the structures that belong to the same developmental stage. In PT studies, using implicational scaling as a metric to measure consistency across individual learners' rank orders is standard practice, as it can account for learner variation within the theorized constraints of PT (Pienemann, 1998a).

Learning Trajectories. For the purpose of examining learning trajectories, we perform a rank

[4] What number of examples that constitutes sufficient evidence varies across languages and studies. For example, while Pienemann (1998b) has initially suggested minimally one occurrence per sample for the syntactic structures as evidence for emergence, Håkansson and Norrby (2010) required two occurrences in their study.

correlation permutation test,[5] inspired by Evanson et al. (2023) and Liu et al. (2021). We rank the PT structures in terms of their acquisition time (the number of steps taken to reach an accuracy above the respective acquisition threshold) and then compute the rank correlation between each pair of the five models and average it. A null distribution is then created by randomly shuffling the ranks in one model per pair and recomputing the average correlation. If the true average correlation is higher than the correlation from the null distribution, the acquisition trajectory is consistent.

4 Results and Discussion

4.1 Performance on SwePT

In addition to the fine-tuned models, we evaluated the pretrained English GPT-2 model (without fine-tuning) as well as the 126M parameter GPT-SW3[6] model. GPT-SW3 was pretrained on 320B tokens of text in Scandinavian languages, mainly Swedish, and thus functions as a skyline. Table 2 displays the final accuracies from the AJT on SwePT of all models. GPT-SW3's average accuracy score 95.38% roughly aligns with the manually calculated precision score of 97.11% (see Appendix A.3).

The accuracies across all structures and fine-tuned models, with the exception of NEG in `GPT-reverse`, are higher than the English pretrained GPT-2 but lower than the GPT-SW3. This indicates that while the fine-tuning was successful, 20M Swedish tokens in the fine-tuning data cannot compare in size to the 320B tokens that GPT-SW3 was trained on and is likely insufficient to reach maximum performance.

Accuracy on PR_a and PR_b remains below chance for all fine-tuned models. GPT-SW3's higher performance on these structures suggests that attractor-sensitive hierarchical generalization in PR_b is weaker than heuristic-based strategies, though this conclusion is tentative given the small number of PR_b minimal pairs.

The low NEG accuracy in `GPT-reverse` is likely due to catastrophic forgetting (McCloskey and Cohen, 1989): NEG appears only during the first 500 time steps, after which post-negations dominate training. The superior overall performance of `GPT-order` supports this explanation, as earlier structures are repeatedly reintroduced later in training, reducing forgetting.

4.2 Acquisition Time

Acquisition time is defined as the checkpoint at which accuracy first exceeds a given threshold. Table 8 in Appendix E displays the acquisition times across models and thresholds. The hypothesized PT implicational order can be inferred from the acquisition times of `GPT-order`, where at least one structure per stage is acquired before or simultaneously as higher-stage structures, with the exception of INQ, which is expected given the input order.

There is a noticeable variability in acquisition times within stages. PLU emerges over 1000 time steps later than SVO and TNS at the 50% and 60% thresholds and never reaches 80%. INV crosses the 50% threshold early, while PR structures are acquired late or not at all. Although SVO and TNS (Stage 2) are generally acquired early, in `GPT-reverse` they emerge after NEG (Stage 5), which is introduced first to `GPT-reverse`. This sensitivity to input order suggests that PT predictions are not robust under curriculum manipulation, consistent with the implicational scaling results. A rank-correlation permutation test shows consistent relative acquisition orders across all models (Table 9, Appendix E), despite differences in absolute timing.[7]

4.3 Implicational Patterns

Table 7 presents collapsed implicational scales using thresholds at 50%, 60% and 80% accuracy.

While the observed order differs slightly between the three scales, all observed patterns deviate from the predicted order as hypothesized by PT. In all scales, higher-stage structures emerge before lower-stage ones; for instance, INQ (Stage 5) precedes PR_a, PR_b, and PLU. There is also significant variability within each scale. The IR (index of reproducibility) coefficients across all three scales are far below the 0.93 scalability threshold (Rickford, 2004). This implies that the observed order is not implicational.

4.4 Learning Trajectories

Figure 1 displays the acquisition trajectories of all linguistic structures tested through the AJT. The

[5] e.g., the Spearman's coefficient of rank correlation, or Spearman's ρ (Gibbons and Chakraborti, 2014).

[6] https://huggingface.co/AI-Sweden-Models/gpt-sw3-126m

[7] Note that that the rank correlation test only measures the rank order at the predefined acquisition thresholds, meaning that the rank order across the entire fine-tuning sequence is not reflected in these results. See Figure 1 for the ordering of the grammatical structures across the entire sequence.

Model	SVO	PLU	TNS	ATT	PR_a	PR_b	INV	NEG	INQ	Avg.
SW3	**98.57%**	**99.58%**	**95.10%**	**94.62%**	**90.61%**	**88.89%**	**93.06%**	**99.01%**	**98.94%**	**95.38%**
GPT-2	57.53%	10.86%	54.75%	26.85%	24.88%	40.74%	51.65%	38.94%	51.06%	39.70%
mixed	91.50%	63.26%	86.90%	74.96%	45.07%	44.44%	67.61%	70.63%	86.17%	70.06%
mixed_2	90.75%	64.93%	86.50%	75.00%	45.54%	48.15%	66.80%	71.95%	85.11%	70.52%
order	90.51%	63.26%	84.65%	74.07%	44.13%	44.44%	72.30%	95.38%	88.30%	73.00%
reverse	91.07%	62.21%	87.95%	50.13%	32.39%	44.44%	56.95%	28.71%	85.11%	59.89%

Table 2: Results from evaluating the last checkpoints of all fine-tuned models, the Swedish GPT-SW3 model and the base English GPT-2 model on SwePT.

results from the `GPT-mixed_2`, which follows a very similar pattern to the `GPT-mixed` model, is presented in Figure 3 in Appendix E.1.

`GPT-mixed`, which was trained without curriculum learning, displays the most consistent trajectory, with higher average accuracies compared to the two curriculum models. With the exception of PR_b, which should be considered an outlier due to its minimal dataset size, the acquisition of all structures follow a visibly parallel acquisition pattern in an order that defies the predictions of PT. NEG and INQ which are hypothesized to be the most difficult for the model to learn since they require the processing procedure at the 5th developmental stage, quickly reach above the 60% acquisition threshold before structures from the 4th and 3rd stages. The PR subsets never reach above the 60% threshold, but rather decrease in accuracy throughout training. Notably, in all models SVO, TNS, INQ and INV have already reached an above-chance accuracy at the first checkpoint.

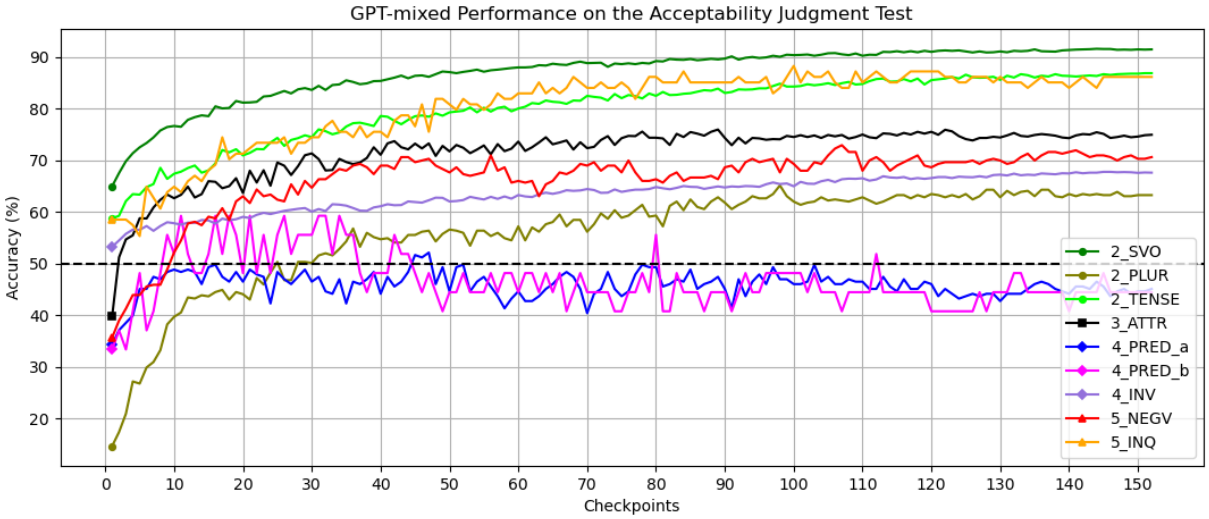
In the curriculum models (Figures 1b and 1c), accuracy closely tracks the distribution of structures in the training data (Table 4). NEG illustrates this clearly: in `GPT-reverse`, it exceeds 90% accuracy during early Stage-5 training, then steadily declines to 30% by the final checkpoint. The reverse pattern appears in `GPT-order`. ATT shows a similar curriculum-dependent rise, with accuracy increasing sharply when its corresponding stage is introduced.

4.5 Curriculum Learning Effects

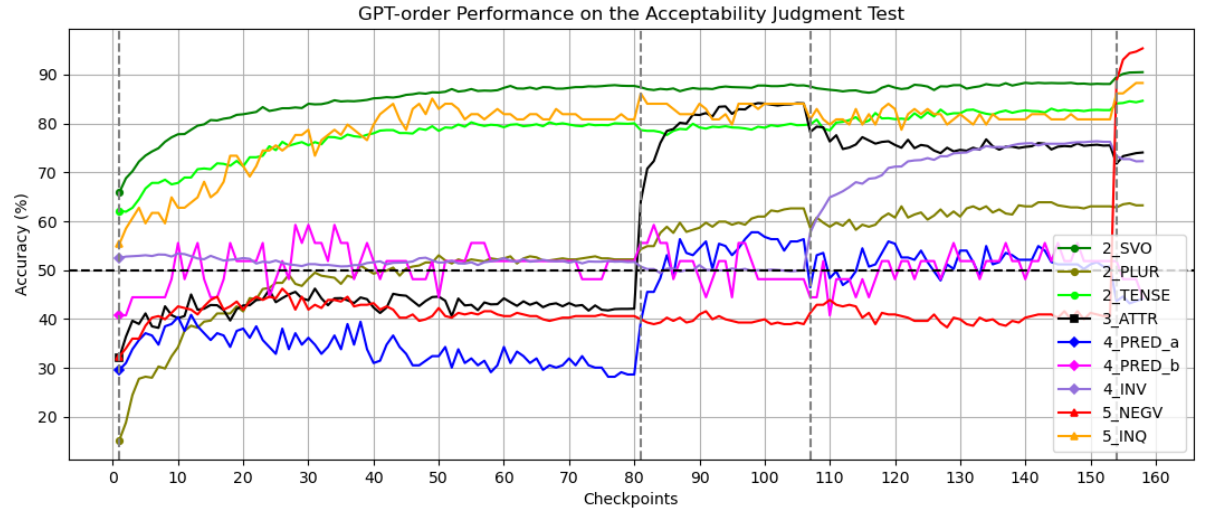
Although the acquisition order contradicts PT, trajectories are systematic and align with prior findings (see Section 2.1) that LLMs behave similarly to bag-of-words models during initial training stages (Choshen et al., 2021). To examine this pattern of unigram statistics, we calculated the distribution of the contrasting morphological features in each minimal pair of SwePT, by searching the dependency trees using simple rule-based scripts. The distribution (see Appendix A.2) reveals that the models are seemingly favoring the sentence of each minimal pair that contains the more frequent form of the target structure. The initial 10% accuracy of the PLU structure roughly correlates with the 80-20 ratio between singular (lemma) and plural nouns in the training data. A similar pattern is detected for the initial below-chance accuracy of ATT, since attributive adjectives in common form (lemma) are in majority in the training data. For TNS, the initial *high* accuracy aligns with the observation that tensed verbs are more frequent than their respective infinitive forms (lemma) in the training data. As training progresses, the accuracy curves of PLU and ATT in all models rise quickly, suggesting that the classifications deviate more and more from the observed unigram distribution. This is in line with previous research indicating that LLMs in later learning stages start to resemble n-gram models that are sensitive to word order, and eventually start relying more on structural cues in context (Saphra and Lopez, 2018; Choshen et al., 2021).

Interestingly, the accuracy on the PR subsets stays around chance-level in the mixed model, while performance on PR_a fluctuates in predictable patterns for `GPT-order` and `GPT-reverse` with regards to the training data. The minimal pairs for PR_a were constructed using the same principles as for ATT, where neuter gendered or plural form of the adjective is contrasted to the common, singular form of the same adjective in the ungrammatical sentence. This implies that the grammatical sentence in the PR_a subset will be subject to the same frequency-based bias as in ATT, as previously discussed, thus favoring the ungrammatical version of each pair. PR_b, while smaller and manually constructed without this bias, also shows no improvement, suggesting that the models have not reached training stages where global hierarchical cues dominate, such as agreement beyond the NP (Stage 4 in PT).

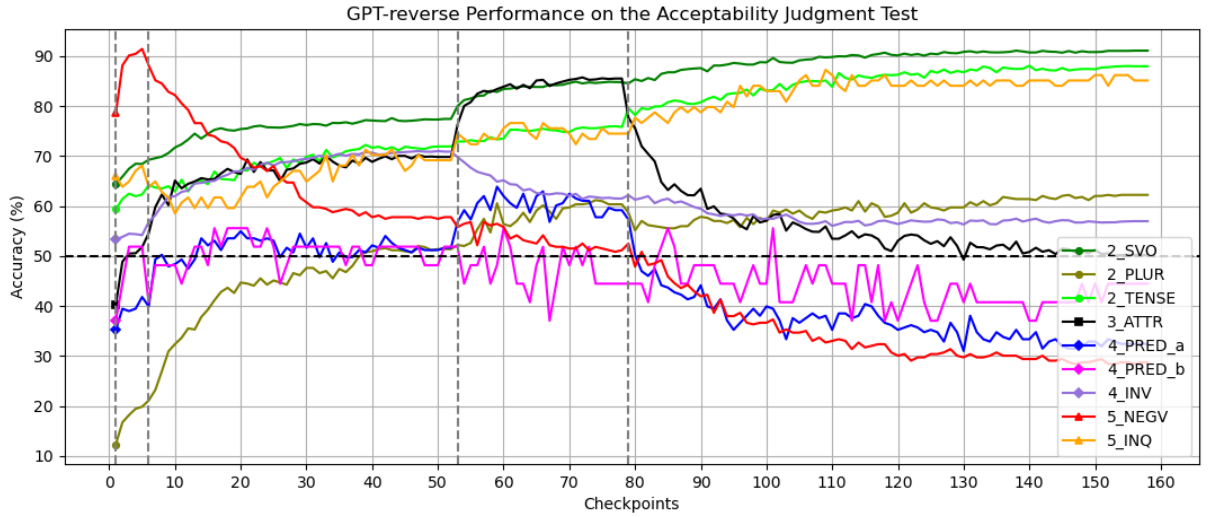
INQ consistently shows high accuracy across all



(a) Results from the AJT on all linguistic structures across checkpoints for the `GPT-mixed` model trained on all four data subsets concatenated and shuffled.



(b) Results from the AJT on all linguistic structures across checkpoints for `GPT-order`, trained on a curriculum with increasing difficulty as hypothesized by PT, in the order Stage-2–5.



(c) Results from the AJT on all linguistic structures across checkpoints for `GPT-reverse`, trained on a curriculum with decreasing difficulty as hypothesized by PT, in the order Stage-5–2.

Figure 1: AJT results across checkpoints for GPT-2 fine-tuned under three different conditions. The vertical dashed lines in `GPT-reverse` and `GPT-order` mark where training commences on data from a new Stage subset.

models despite its hypothesized difficulty in PT. This can be explained by the fact that interrogative subclauses share syntactic structure with declarative subclauses and are likely found across all stages in the training data. As a result, models can generalize INQ word order early. The similarity between INQ and Stage-2 structures (SVO, TNS) suggests that GPT-2 relies on syntactic regularities rather than semantic overgeneralization. Given evidence that semantic information is learned later than syntax (Blevins et al., 2022; Saphra and Lopez, 2018), a single epoch of fine-tuning may be insufficient for models to adopt human-like semantic strategies, even though simpler n-gram-based heuristics prove effective for INQ.

One could speculate that the window of time during which the models were evaluated in this study only reflects performance from the 2nd developmental stage, which would not reject the hypothesis that the models enter further developmental stages later in training. The observed high accuracy on the structures from the 3rd, 4th and 5th developmental stages could then be explained through surface-level heuristics, which could be argued to require only the lemma and category procedures available at the 1st and 2nd developmental stages.

It is also possible that GPT-2 does not possess the same processing constraints as humans do, that such constraints are induced later in training, that different learning strategies are applied or that PT truly cannot be tested through receptive skills, as discussed by Pienemann (1998b) and Dyson and Håkansson (2017). Alternatively, if the relevant distinction is not between production and reception but rather between implicit and explicit knowledge, as argued by Ellis (2008), it may be the case that the "cognitive processes" behind GPT-2's next-word prediction, guiding its classifications and text generation, align more with the "explicit knowledge" utilized by humans during e.g. an untimed AJT than with "implicit knowledge" used for timed AJT and speech production.

5 Conclusion

The present study examined whether GPT-2's acquisition order of Swedish grammatical structures follows the order sequence as stipulated by Processability Theory, and to what extent this acquisition order is affected by the input sequencing of the training data (i.e., different curricula). The results indicated that while the observed acquisition order was found to be robust to the order sequencing of the training data as measured with rank correlation tests at the thresholds of acquisition defined in this study, the acquisition order of the fine-tuned models did not align with the implicational order sequence as hypothesized by PT. Observations of the performance on the AJT and the frequency distribution of the contrasting features in the minimal pairs suggested that the performance can largely be explained by unigram and n-gram heuristics. These findings suggest that the grammatical development predicted by PT does not naturally emerge from next-word prediction objectives. These results should be interpreted with caution, however, due to inherent incompatibilities between the PT framework and the methodology required for testing grammatical receptive skills with AJT.

Limitations and Future Work

Adapting the emergence criterion. As previously addressed by authors of PT studies focusing on receptive skills, the emergence criterion cannot be seamlessly adapted to align with the evaluation of the AJT. A fundamental conceptual distinction that this study fails to resolve remains: the fact that accuracy and emergence reflect separate aspects of acquisition. Since the accuracy rates of different structures develop with different gradients, the inferred acquisition order is sensitive to the predefined acquisition threshold, which reduces reliability. While the plotted learning trajectories in this study offer transparency regarding this aspect, by modeling the change in accuracy across time, the gradient property inherently does not align with the categorical logic of the emergence criterion. This has a significant impact on the validity of the results as evaluated within the framework of PT and requires further addressing.

Furthermore, evidence of rule application in its obligatory context does not only encompass grammatically correct target-like forms, but any application that can demonstrate that the learner is able to process that grammatical rule[8]. By limiting the evaluation to the binary classification of a minimal pair, valuable information from non-target con-

[8]For example, the past tensed form of the Swedish irregular verb "gå" (*walk*) is "gick", while a majority of regular verbs are realized in past tense with the *-de* suffix. Thus, the interlanguage form "*gådde" often emerge in the speech output of Swedish L2 learners that have access to the processing procedure on the 2nd developmental stage, serving as evidence for access to the procedure despite its incorrect surface form (Flyman Mattsson, 2022).

structions may be lost (see e.g. Schönström, 2014).

Future work could avoid these methodological issues by focusing on the language *production* of LLMs, where the emergence criterion could be implemented without the need for adaptation to accuracy scores, and potential "interlanguage" constructions of the target grammatical rule could be taken into account, offering insights beyond the limitations of minimal pairs.

Minimal pair dataset generation. The minimal pairs of SwePT were identified and altered on the basis of a single metric of complexity: the presupposed processing constraints of its target linguistic structure. Each sentence is thus assumed to be equally as complex, regardless of aspects such as token frequency and sentence length that are confounding factors to the acceptability of a sentence as predicted by probability. While sentence length does not impact the difference in score between sentences of a single minimal pair that are equal in length, it may impact the overall processability and make the acceptability scores less reliable, as may differences in parse-depth across pairs.

Furthermore, the identification constraints for the morphological structures could allow more variety of obligatory context to avoid tying the performance on the AJT to a single heuristic. For example, the TNS subset may elicit different performance on the AJT with more variation of verb forms across the pairs, and with the obligatory context for a specific tense based on semantic or syntactic cues rather than context-independent occurrences. Within pairs, the difference in semantic plausibility after altering the grammatical sentence is not accounted for. Manual refinement of the scripts where length, token frequencies and an extended set of obligatory contexts are controlled for may increase the interpretability of the AJT results.

The observed number of false positives could also be reduced by utilizing native-speaker crowdsourcing and/or a Swedish LM such as GPT-SW3 for evaluation of the minimal pairs.

Parsing evaluation. While the quality of large common crawl datasets such as OSCAR is hard to control for, in future work, it is recommended to evaluate the fine-tuning data parsing process more rigorously in order to quantify the noise. KL divergence may be a better choice than the chi-square test when quantifying how much the parsed fine-tuning data diverges from the gold-standard distributions. Processing in earlier steps, including the splitting of sentences before parsing, should also be evaluated systematically to ensure its precision and avoid propagating errors further down the pipeline.

Additional models. With additional computing resources, including additional models beside GPT-2 would increase the relevance of the study. In order to offer additional insights on transfer effects and the modeling of first and second language acquisition, a Swedish model such as GPT-SW3 could be pretrained and evaluated, providing a comparison between the acquisition order of a fine-tuned "L2 learner" and itself as an "L1 learner".

References

Gerry TM Altmann and Yuki Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3):247–264.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2022. Analyzing the mono-and cross-lingual pretraining dynamics of multilingual language models. *arXiv preprint arXiv:2205.11758*.

Aafke Buyl and Alex Housen. 2015. Developmental stages in receptive grammar acquisition: A processability theory account. *Second Language Research*, 31(4):523–550.

Cheng-Han Chiang, Sung-Feng Huang, and Hung-yi Lee. 2020. Pretrained language model embryology: The birth of albert. *arXiv preprint arXiv:2010.02480*.

Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. 2021. The grammar-learning trajectories of neural language models. *arXiv preprint arXiv:2109.06096*.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Bronwen Patricia Dyson and Gisela Håkansson. 2017. *Understanding second language processing: A focus on Processability Theory*, volume 4. John Benjamins Publishing Company.

Rod Ellis. 2008. Investigating grammatical difficulty in second language learning: Implications for second language acquisition research and language testing. *International journal of applied linguistics*, 18(1):4–22.

Linnea Evanson, Yair Lakretz, and Jean-Rémi King. 2023. Language acquisition: do children and language models follow similar learning stages? *arXiv preprint arXiv:2306.03586*.

Anna Flyman Mattsson. 2022. Rethinking textbook grammar introduction. *Instructed Second Language Acquisition*, 6(2):196–218.

Jean Dickinson Gibbons and Subhabrata Chakraborti. 2014. *Nonparametric statistical inference: revised and expanded*. CRC press.

Gisela Håkansson and Catrin Norrby. 2010. Environmental influence on language acquisition: Comparing second and foreign language acquisition of swedish. *Language learning*, 60(3):628–650.

Fredrik Heinat. 2012. Finiteness in swedish. *Working papers in Scandinavian syntax*, 90:81–110.

Olle Josephson. 2020. Grammatik, ord, texttyper: Svenska med fokus på form.

Satomi Kawaguchi. 2008. Argument structure and syntactic development in japanese as a second language. In *Cross-linguistic aspects of processability theory*, pages 253–298. John Benjamins Publishing Company.

Dagmar Keatinge and Jörg-U Keßler. 2009. The acquisition of the passive voice in l2 english: Perception and production. *Research in second language acquisition: Empirical evidence across languages*, pages 67–92.

Tatsuki Kuribayashi, Yohei Oseki, Souhaib Ben Taieb, Kentaro Inui, and Timothy Baldwin. 2025. Large language models are human-like internally. *arXiv preprint arXiv:2502.01615*.

Olivia La Fiandra, Nathalie Fernandez Echeverri, Patrick Shafto, and Naomi Feldman. 2025. Large language models and children have different learning trajectories in determiner acquisition. In *Proceedings of the First BabyLM Workshop*, pages 100–108.

Willem Levelt. 1989. Speaking-from intention to articulation. *A Bradford book*.

Leo Z Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A Smith. 2021. Probing across time: What does roberta know and when? *arXiv preprint arXiv:2104.07885*.

Fethi Mansouri. 2008. Agreement morphology in arabic as a second language: Typological features and their processing implications. In *Cross-linguistic aspects of Processability Theory*, pages 117–153. John Benjamins Publishing Company.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

R Thomas McCoy, Robert Frank, and Tal Linzen. 2018. Revisiting the poverty of the stimulus: Hierarchical generalization without a hierarchical bias in recurrent neural networks. *arXiv preprint arXiv:1802.09091*.

Catrin Norrby and Gisela Håkansson. 2007. The interaction of complexity and grammatical processability: The case of swedish as a foreign language. *Iral-international Review of Applied Linguistics in Language Teaching - IRAL-INT REV APPL LINGUIST*, 45:45–68.

Manfred Pienemann. 1998a. Developmental dynamics in l1 and l2 acquisition: Processability theory and generative entrenchment. *Bilingualism: Language and cognition*, 1(1):1–20.

Manfred Pienemann. 1998b. *Language processing and second language development: Processability theory*, volume 15. John Benjamins Publishing.

Manfred Pienemann. 2005. *Cross-linguistic aspects of processability theory*. John benjamins publishing company.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

John R Rickford. 2004. Implicational scales. *The handbook of language variation and change*, pages 142–167.

Naomi Saphra and Adam Lopez. 2018. Understanding learning dynamics of language models with svcca. *arXiv preprint arXiv:1811.00225*.

Krister Schönström. 2014. Visual acquisition of swedish in deaf children: An l2 processability approach. *Linguistic approaches to bilingualism*, 4(1):61–88.

Patti Spinner. 2013. Language production and reception: A processability theory study. *Language Learning*, 63(4):704–739.

Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. 2024. Rublimp: Russian benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2406.19232*.

Elena Volodina, Yousuf Ali Mohammed, Aleksandrs Berdičevskis, Gerlof Bouma, and Joey Öhman. 2023. Dalaj-ged-a dataset for grammatical error detection tasks on swedish. In *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*, pages 94–101.

Elena Volodina, Yousuf Ali Mohammed, and Julia Klezl. 2021. Dalaj-a dataset for linguistic acceptability judgments for swedish: Format, baseline, sharing. *arXiv preprint arXiv:2105.06681*.

Xiaojing Wang. 2011. *Grammatical development among Chinese L2 learners: From a processability account*. Ph.D. thesis, Newcastle University.

A Warstadt. 2019. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, et al. 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*.

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.

Aditya Yedetore, Tal Linzen, Robert Frank, and R Thomas McCoy. 2023. How poor is the stimulus? evaluating hierarchical generalization in neural networks trained on child-directed speech. *arXiv preprint arXiv:2301.11462*.

A Details on SwePT

A.1 Processing Pipeline Details

The Swedish Talbanken and LinES treebanks from UD were merged into a single CoNLL-U data file before processing. Double citation marks were removed from the sentences due to spacing issues during the conversion into the ungrammatical sentences. All tokens without integer indices (floats representing implicit, omitted words in elliptical structures) were skipped, since these tokens are not explicit in the original sentences. In cases where more than one instance of the target structure was found in the same sentence, only one instance was modified in the ungrammatical sentence.

The pipeline handles whitespace before and after punctuation to ensure that the grammatical and ungrammatical sentences do not differ in more aspects than the target structure. The scripts allow no duplicates and only returns the minimal pair if the grammatical and ungrammatical sentence are not identical.

The following section contains examples and explanations of each grammatical structure of the Swedish PT hierarchy, as well as the criteria for identification of each grammatical structure.

A.1.1 Canonical word order (SVO)

The category procedure at stage 2 does not contain any unification or information exchange between constituents, but enables mapping syntactic categories and functional roles to the lexicon such as subject and predicate, allowing the learner to structure sentences in canonical SVO order. Swedish is a verb-second language, which means that the finite verb is always placed directly after the topicalized constituent in main clauses (and occasionally after selected clause adverbials). Thus, observed *SOV word order is ungrammatical and indicates that the learner has not yet accessed the category procedure. The contrast between SVO and *SOV is illustrated in examples 1 and 2 below.

(1) Jag läser boken
I read.FIN book

'I read the book'

(2) *Jag boken läser
I book.FIN read

The selection of sentences for the SVO subset was made based on the following constraints:

1. The first token of the sentence (or second in case of initial quotation marks) must not be a relative pronoun, an interrogative pronoun, a subjunction or a verb.
2. The sentence must contain a subject which must be a noun, a proper noun, a pronoun or a determiner.
3. The subject must be governed by a root verb.
4. The subject must precede the root.[9]
5. The sentence must contain an object or clausal complement which must be a dependent of the root.

To form the ungrammatical sentence for the minimal pair, all dependents of the subject and object phrases were identified, and the object phrase was positioned in front of the finite verb in the ungrammatical sentence, forming *SOV word order.

A.1.2 Plural (PLU)

Access to the category procedure also enables encoding lexical information, which is necessary for marking plural on nouns and tense on verbs. Plural

[9]This ensures that no sentences with topicalization or subjunctive or imperative root verbs are included.

is normally marked morphologically as a suffix on the noun with *-or, -ar, -r* or *-n* in indefinite form and *-na* or *-en* in definite form, depending on the declension of the noun. While the access to the category procedure in humans' language output can be assessed from all occurrences of plural and non-occurrences of plural in their obligatory contexts, such obligatory contexts must be explicit when using AJT. As an example, the sentence "Elefanterna är här" (*The elephants are here*) and "Elefanten är här" (*The elephant is here*) are equally as grammatical in Swedish, and while the context may be inferred in a speech sample, it cannot be inferred in an AJT where the sentences are isolated. Moreover, while the plural inflection on an attributive adjective modifying a plural noun such as *-a* on "stor" in "De stora elefanterna är här" (*The big elephants are here*) does serve as obligatory context for plural, it cannot be discerned whether recognizing such a context is due to proper processing of plural or of attributive agreement. Thus, plural numerals were selected to form the obligatory context for the plural feature on the noun used in this study, as in example 4 below.

(3) En grå elefant
one gray.SG elephant.SG

'One gray elephant'

(4) Två grå-a elefant-er
two gray-PL elephant-PL

'Two gray elephants'

(5) *Två grå-a elefant
two gray-PL elephant.SG

The selection of sentences for the PLU subset was made based on the following constraints:

1. The sentence must contain a numeric modifier of which the lemma is not "1", "en" or "ett".

2. The sentence must contain a plural noun that governs the numeral.

3. The noun must not be marked with the genitive case.[10]

4. The lowercased form of the noun must not be the same as the noun's lemma.[11]

5. The noun must not be an abbreviation.[12]

Note that the constraints in the script allow plural nouns that occur in definite form, e.g. "de fem nedersta trappstegen" (*the five lowest steps*). In the PT hierarchy, definite form and plural are found on the same developmental stage.

To form the ungrammatical version of the minimal pair, the noun in each sentence was modified into its lemma, i.e. into its singular form.

A.1.3 Tense (TNS)

In Swedish, tense is marked morphologically with inflection or suffixes. While the difference in acceptability when using one tense above another (e.g. "Jag ser[PRS] ett träd" (*I see a tree*) and "Jag såg[PST] ett träd" (*I saw a tree*)) might be inferrable from the context of the utterance, such context is not available in AJT. Thus, in this study the obligatory context of tense is simply identified in all occurrences of tensed verbs, with the verb converted into infinitive in its ungrammatical equivalent. This is illustrated in examples 6 and 7 below.

(6) Jag såg ett träd
I see.PST a tree

'I saw a tree'

(7) Jag se ett träd
I see.INF a tree

The selection of sentences for the TNS subset was made based on the following constraints:

1. The sentence must contain a tensed verb or an auxiliary.

2. The verb must be in active form.

3. The verb form must be distinct from its lemma.

To form the ungrammatical version of the minimal pair, the verb in each sentence was modified into its lemma, which corresponds to its infinitive form.

[10] Genitive in Swedish is marked with an "s" at the tail of the word and does not change other inflections of the noun. Thus, this is not an obligatory constraint, but simply serves to decrease the number of matches and simplify the subset.

[11] This excludes cases where there is no distinction between the singular and plural form, e.g. "ett hus, två hus" (*one house, two houses*).

[12] This excludes cases such as "kr" (the abbreviation of the Swedish currency *kronor*) which makes no distinction between the singular and plural but where the form differs from the lemma "kronor".

A.1.4 Definiteness (N/A)

The processing required for the morphological marking of definiteness on nouns is also enabled on the second developmental stage of the PT hierarchy. Since three other grammatical structures are already included to represent the 2nd developmental stage in the current study however, definite form as a category is excluded from analysis in this study due to time limitations.

A.1.5 Non-inversion with topicalization (N/A)

The phrasal procedure at stage 3 enables unification of features within the phrase, and thus gives access to the processing of attributive agreement and topicalizing constituents. However, the process necessary for obligatory inversion of the verb after a topicalized constituent is not accessible until stage 4. This implicates that the language output containing topicalization indicating that a learner has reached stage 3 is ungrammatical by default. Thus, the model cannot be tested on the *XSV structure of this stage, which is why this structure is excluded in the present study.

A.1.6 Attributive agreement (ATT)

Attributive agreement entails that features of gender (common and neuter), number (singular and plural) and definiteness (indefinite and definite) are unified within the noun phrase and marked on both article, noun and adjective, and thus requires access to the phrasal procedure. The example below showcases this exchange of features. Observe that the plural form of the adjective is identical to its definite singular form. The present study uses the common form of the adjective in the obligatory context of neuter or plural agreement as evidence for ungrammaticality. This entails that attributive agreement with common singular nouns are not identified as examples of this structure.

(8) Den stor-a hund-en
DET.COM.SG.DEF big-SG.DEF dog.SG-DEF.COM

'The big dog'

(9) Många stor-a hund-ar
many big-PL.IND dog-PL.COM.IND

'Many big dogs'

(10) *Många stor hund-ar
many big.SG.COM.IND dog-PL.COM.IND

The selection of sentences for the ATT subset was made based on the following constraints:

1. The sentence must contain an adjectival modifier.
2. The adjective must be in positive and indefinite form.
3. If in singular form, the adjective must not be in common/utrum gender, and the form must be distinct from its lemma.

To form the ungrammatical version of the minimal pair, the adjective in each sentence was modified into its lemma (and capitalized when the form was capitalized), which corresponds to its common and singular form.

A.1.7 Predicative agreement (PR_a and PR_b)

Grammatical information of number and gender is also exchanged interphrasally between the noun in the NP and the adjective in the VP. Mismatching the form of the predicative adjective with the form of the noun it governs is sufficient as counter-evidence for acquisition of the predicative agreement structure. Example 11 and 12 illustrate a grammatical and ungrammatical sentence with regards to predicative agreement, including an attractor phrase within brackets. Including an attractor noun may reveal patterns of the model's generalization strategies based on whether or not it marks the adjective with the grammatical information of the attractor in favor of the subject due to their linear proximity.

(11) Hund-ar-na [på gård-en] är stor-a
dog-PL-DEF.COM on yard.SG-DEF.COM be big-PL

'The dogs on the yard are big'

(12) *Hund-ar-na [på gård-en] är
dog-PL-DEF.COM on yard.SG-DEF.COM be
stor
big.SG.COM

The PR subset is separated into PR_a and PR_b, where the former is created by altering the predicative adjective in the ungrammatical sentence to correspond to its lemma, and the latter includes an attractor noun that differs from the subject in number or gender, and that is linearly closer to the predicative adjective than the subject.

The sentences for PR_a were selected based on the following constraints:

1. The subject must be a noun.
2. The sentence must contain a copula.

3. The predicate agreement must occur between the subject and an adjective in positive form that governs the subject.

4. The subject must precede the copula.

5. The lowercase form of the adjective must differ from its lemma.

For PR_b, the sentence should contain a second noun that operates as an attractor. The following constraints were applied:

1. The noun subject must be governed by a root adjective,

2. The sentence must contain a second noun (attractor) which is governed by the subject,

3. The attractor must have the noun modifier relation.

In order to ensure that the adjective explicitly agrees with the subject and can be modified into a form that agrees with the attractor, sentences are excluded if

1. the adjective inflection makes no distinction between neither gender nor grammatical number,[13]

2. the adjective inflection makes no distinction between gender in its form,[14] and the attractor is in singular,

3. the attractor and the subject are marked with the same gender and are both marked with singular,

4. the subject and the attractor are both marked with plural,

5. the subject and attractor differ in gender but the adjective does not have a singular lemma.

After applying the constraints, only 28 sentences were elicited. Due to this scarce number, the duplicated sentences were *manually* altered into their ungrammatical form by modifying the form of the adjective to modify the attractor. In some cases where the attractor consisted of multiple nouns, the noun phrase was simplified to contain only the first noun.Since the constraints in the script for PR_a are also in place for PR_b, there was an overlap of sentences in both subsets after processing. Thus, as a last step, the sentences from PR_a that also occurred in PR_a were removed from PR_a.

It is important to note that the script does not take into account the occasions where the predicative agreement is governed by semantics rather than grammar. One systematized example of this is the phenomenon of the singular neuter form of an adjective being used to describe an abstract noun, regardless of the grammatical gender or number of that noun (e.g., "skatteberäkning[COM] kan vara jobbigt[NEU] att utföra" (*Tax calculations can be difficult to perform*, or "En avromantisering[COM] av äktenskapet är nödvändigt[NEU] för kvinnans egen skull" (*A deromantization of marriage is neccessary for the woman's own benefit*)). Nouns that are conceptually interpreted as an entity or situation are generally referred to by the general neuter determiner "det" (*it*), which triggers the acceptability of the neuter agreement on the adjective. Other examples where multiple adjective markings may be acceptable, albeit not grammatical, include the singular vs. plural agreement with nouns representing groups of people, such as "Nämnden[COM, SING] var inte beredda[PLU] att ta ett beslut i frågan" (*The committee was not prepared to make a decision on the matter*).

The abovementioned phenomena serve as examples of the fact that grammaticality and acceptability are not equivalent concepts. Sentences that violate these grammatical aspects were manually removed in the PR_b subset, but may occur in the PR_a subset.

A.1.8 Inversion after topicalization (INV)

Through the procedure on stage 4, interphrasal information can be exchanged, allowing the learner to properly use inversion of the finite verb when topicalizing constituents (example 13b below). Swedish is a verb-second language, which means that the finite verb must always be placed directly after the topicalized constituent (XVS). Before the interphrasal procedure stage is acquired, learners will use topicalization without the obligatory inversion of the verb (*XSV), as illustrated in the ungrammatical sentence in example 15 below.

(13) Jag ska läsa boken imorgon
I will.FIN read book tomorrow

'I will read the book tomorrow'

[13] Examples are the adjectives "skrämmande" or "bra", that can modify nouns of any grammatical gender or number.

[14] An example is the adjective "indiskret", which retains its form when modifying neuter nouns.

(14) Imorgon ska jag läsa boken
tomorrow will.FIN I read book

'Tomorrow I will read the book'

(15) *Imorgon jag ska läsa boken
tomorrow I will.FIN read book

The selection of sentences for the INV subset was made based on the following constraints:

1. The first constituent (after any punctuation or conjunctions) before the finite verb must not be a subject (passive or active), expletive, interrogative pronoun or imperative verb.

2. The sentence must contain a subject (active or passive) or expletive.

The script operates by identifying the root verb or auxiliary dependent of the root verb and identifying all constituents that precede this verb, as well as all the subject dependents. The position of the subject phrase is then switched with the finite verb/auxiliary to form the ungrammatical sentence of the minimal pair.

Observe that the adverb "kanske" (*maybe*) in the initial position in the main clause sometimes is found with a directly succeeding subject.[15] In other words, it is realized grammatically as a conjunction rather than an adverb. Although it is tempting to filter out such occurrences, all scripts operate based on grammar principles, not acceptability principles, and an exception should not be made here.

A.1.9 Pre-verbal negation in subclauses (NEG)

The subordinate clause procedure of stage 5 does not contain any unification of features, rather it has as its prerequisite the acquisition of all word order constraints of the main clause. In Swedish, the access to the procedure of this developmental stage is revealed by the L2 learner placing the negation *inte* and other clausal adverbs in front of the finite verb in a subclause. The ungrammatical equivalent would be placing the negation after the finite verb, which is grammatical only in main clause syntax. The subject may precede or follow the negation (see "det" in examples 16 and 17 below). Before this process is accessed, learners generalize the SVneg word order from main clauses to subclauses, resulting in ungrammatical sequences such as example 18.

[15]Compare the sentences "Kanske är jag hungrig" and "Kanske jag är hungrig", which are both acceptable in Swedish (see e.g., Heinat, 2012).

(16) Jag går inte om det inte är kul
I go.FIN not if it not is.FIN fun

'I am not going if it is not fun'

(17) Jag går inte om inte det är kul
I go.FIN not if not it is.FIN fun

'I am not going if it is not fun'

(18) *Jag går inte om det är inte kul
I go.FIN not if it is.FIN not fun

The python script used for selecting and processing sentences for the NEG subset used the following constraints:

1. The sentence must contain a negation with the lemma "inte".

2. The negation head must either be a clausal subject, a clausal complement, an adverbial clause modifier, a clausal modifier of a noun or a relative clause modifier.

3. The negation must precede its head.

4. The negation must not be topicalized.

The script functions by identifying the embedded negation, the embedded verb (and its dependent auxiliary, if applicable) and the embedded subject. The duplicated sentence is then altered into its ungrammatical form by switching the positions of the negation and the finite verb/auxiliary. In cases where the subject succeeds the negation rather than preceding it, the subject is also moved in front of the verb in order to form canonical SVneg(O) word order.

It should be noted that the script also allows for sentences where the precedes the subject in a dependent clause, such as "kostnaderna" in the following Talbanken sentence: "Där kan hyrorna i stort sett inte ändras såvida inte kostnaderna[SUBJ] ökar[...]." (*There, the rents can't be changed much unless the costs[SUBJ] increase.*) This word order results in an ungrammatical equivalent in the subset where the verb precedes the subject ("såvida ökar kostnaderna inte"), i.e. XVSneg main clause word order. Although using a subjunction as a topicalized constituent is not grammatical, with a context window that excludes the first constituent "såvida", the VSneg sequence is grammatical.

It is common that multiple clausal adverbials (such as "inte", "alltid", "fortfarande" (*not, always, still*)) are placed in juxtaposition in a sentence. The position of the negation "inte" in relation to other

clausal adverbials can differ. In some cases, this entails that the clausal adverbials will be separated from each other in the ungrammatical sentence, e.g. in "[...]eftersom jag bara inte har[...]" –> "[...]eftersom jag bara har inte[...]" (*[...]because I just don't have[...]*). When the negation precedes the second clause adverbial however, it results in another word order in the ungrammatical sentence. Ponder the Talbanken sentence "Det är ett jobb som inte[neg] bara kräver[VERB] en eller två föräldrar utan insatser från så många olika håll[...]." (*It's a job that doesn't just require one or two parents, but input from so many different sides.*) Observe that the swapping of the negation and verb in this sentence results in an ungrammatical sentence where the negation doesn't immediately follow the verb. For this particular sentence, the change in word order ("som kräver bara inte en eller två föräldrar, utan[...]") results in an arguably acceptable interpretation of the dependencies, where "bara" (*just*) is related to the noun phrase "en eller två föräldrar" (*one or two parents*) rather than the verb "kräver" (*demands*). It is possible that such cases, if present in the training data, may confuse the model in its acceptability judgments.

A.1.10 Non-inversion in indirect questions (INQ)

Canceling of inversion after question words in interrogative clauses is also accessed on the 5th and final developmental stage. As in English, in Swedish, direct and indirect questions have different word orders. While the verb precedes the subject in direct questions,[16] in indirect questions the subject precedes the verb in the interrogative subclause.[17] Before entering the 5th developmental stage of the PT hierarchy, L2 learners tend to overgeneralize the word order of direct questions to subordinate interrogative clauses in indirect questions, as illustrated in the examples 20 and 22 below.

(19) Jag undrar vad hon inte har gjort
I wonder what she not have.FIN do

'I wonder what she hasn't done'

(20) *Jag undrar vad har hon inte gjort
I wonder what have.FIN she not do

(21) Jag undrar om hon kommer
I wonder if she come.FIN

'I wonder if she's coming'

(22) *Jag undrar om kommer hon
I wonder if come.FIN she

Interrogative subclauses are not grammatically but semantically distinguished from regular relative subclauses. The lemma of the matrix verb is an indicator of the nature of the subclause, where verbs describing inquisitive and cognitive processes such as "undra", "fråga", "fundera", "undersöka", "veta", "gissa", "förklara", "diskutera" and "beskriva" (*wonder, ask, ponder, examine, know, guess, explain, discuss, describe*) are commonly found in the matrix clause (Josephson, 2020). Commonly, the embedded subclause verb (often the head of the question word) functions as a clausal complement.

The python script used for selecting and processing sentences for the INQ subset applied the following constraints:

1. The sentence must contain a matrix verb whose lemma corresponds to "undra", "fråga", "fundera", "undersöka", "veta", "gissa", "förklara", "diskutera" or "beskriva".

2. The sentence must include either a question word or the lemma "om" (*if*) or "huruvida" (*whether*) with the marker relation.

3. The question word must not be the subject.[18]

4. The sentence must contain an embedded verb that governs the question word or marker, and if applicable an auxiliary that is governed by such a verb.[19]

5. If the conjunction has the lemma "om" or "huruvida", the embedded verb must have the clausal complement relation.[20]

6. The embedded clause must contain a nominal subject or an expletive which must not be a relative pronoun.

[16] e.g. "Vad äter du?" (*lit. What eat you (What are you eating)?*)) or "Äter du?" (*lit. Eat you (do you eat)?*)

[17] e.g. "Jag undrar vad/om du äter" (*lit. I wonder what/if you eat*)

[18] as in e.g. "Jag undrar vem som kommer." While this is a valid indirect question, the question word must be separate from the subject in order to generate the ungrammatical equivalent.

[19] This excludes indirect questions where the question word is a dependent of a noun, e.g. in "Jag undrar vilken bok han läser". Since such examples are in minority, and already excluded if containing a subject relative pronoun (e.g. "Jag undrar vilken bok *som* är bra" (*I wonder which book* that *is good*), this constraint was applied in favor of simplicity in the identification of the embedded verb.

[20] This constraint separates indirect questions from conditional clauses.

The script functions by identifying the embedded verb or auxiliary dependent of the embedded verb, as well as identifying the subject phrase in the embedded clause. The position of the verb or auxiliary is then switched with the subject to form the ungrammatical sentence in the minimal pair. If the embedded clause contains a negation that precedes the subject,[21] the negation and finite verb will also swap positions.

A.2 Distribution of Morphological Forms in SwePT

72% of all attributive adjectives in the training data were in common (lemma) form, 27% in neuter and 0.92% in invariant form (gender-agnostic). Among all predicative adjectives, 65% were in common (lemma) form, 35% in neuter and 0.12% in invariant form. Relevant to the PLU subset, 79% of non-genitive nouns (not counting abbreviations) were in singular (lemma) form, and 21% in plural form. Relevant to the TNS subset, 28% of verbs were in infinitive (lemma) form, and 72% in tensed form.

A.3 Evaluation of SwePT

The minimal pair generation was evaluated manually, identifying the positive predictive value (precision) of the minimal pair generation process. 50 random minimal pairs from each subset of SwePT were examined, of which 25 pairs originated from the LinES corpus and 25 pairs from the Talbanken corpus. The false positives, in terms of the number of pairs containing incorrectly identified grammatical structures or incorrect generation of the ungrammatical sentence, were counted. The error rate per minimal pair subset was then calculated with a 95% Wilson Score Confidence Interval in order to account for the small sample size. The minimal pairs were found to have an average error rate of 2.89%, which corresponds to a precision score of 97.11%. The false positives and the corresponding error rates per minimal pair subset are presented in Table 3. Observe that the PR_b subset does not contain any false positives by default, since the ungrammatical sentences were manually generated.

[21] e.g. "Man kan fråga sig om [inte]NEG [detta antagande]SUBJ är felaktigt" (*One may wonder whether (if not) this assumption is incorrect)*

Subset	SVO	PLU	TNS	ATT	PR_a
FP	2/50	2/50	0/50	2/50	3/50
Error rate	4%	4%	0%	4%	6%
Subset	PR_b	INV	NEG	INQ	Avg.
FP	0/50	0/50	2/50	2/50	
Error rate	0%	0%	4%	4%	2.89%

Table 3: Error rates per subset in SwePT. 50 randomly extracted minimal pairs from each subset were examined manually. The false positives (FP) were counted and the error rate was calculated with a 95% Wilson Score Confidence Interval. The average FP score is 2.89%, which corresponds to a precision score of 97.11%.

B Fine-tuning Details

B.1 Fine-tuning Data

Structure	Stage-2	Stage-3	Stage-4	Stage-5	Total
INQ				1 362	1 362
NEG				10 645	10 645
PR_b			51	1	52
PR_a			7 855	217	8 072
INV			111 195	3 156	114 351
ATT		64 569	26 289	2 848	93 706
TNS	194 702	49 118	110 559	11 833	366 212
PLU	13 329	4 001	6 992	397	24 719
SVO	47 764	15 241	15 604	3 004	81 613
N/A	55 268	17 878	31 982	3 295	108 423
n of sents.	254 875	82 447	147 490	15 188	500 000

Table 4: Distribution of the linguistic structure labels of sentences in all four training subsets (stages 2-5).

B.2 Preprocessing of Fine-tuning Data

We used Stanza with the `tokenize`, `pos`, `lemma` and `depparse` tools in our pipeline, with batches of 50 on one GPU. The dataset was first partitioned into ten separate files for parallel processing. The sentences were processed individually in a separate function to minimize the loss of data. Web addresses were also removed using regex matching. Due to memory allocation limits during parsing, the dataset was separated into chunks after timeouts and processed separately based on indexing, after which the files were concatenated into a single CoNLL-U file. Due to unforeseen issues during this process, some of the data was skipped unintentionally. This should be taken into account if replicating this study. Through this process, the dataset was reduced to 25,758,263 sentences/examples, 478,895,789 words and 1,021,799,273 tokens (after truncating with max_length=512).

A Chi-square test shows that the parsed OSCAR data differs significantly from the two gold-standard corpora, in that the syntactic categories (SVO, INV, NEG, INQ) are consistently more frequent in the evaluation data, compared to in the

training data, while the opposite relationship is observed among the morphological structures (ATT, TNS, PLU, PR). See Table 5 and Figure 2 for a visualization of the distribution. Stanza has a reported 87.85 labeled attachment score (LAS)[22] evaluated on the Talbanken treebank. While this is considered a high score for Swedish dependency parsing, in comparison with the performance of other publicly available parsers for Swedish, it does leave room for improvement. It is likely that the observed difference in populations is attributed to error propagation from earlier stages of the processing pipeline, or to natural domain differences between the corpora. In comparison to the manually annotated and corrected Talbanken and LinES, OSCAR as a common crawl corpus is expected to contain noisy data and strings that are independent from grammatical structure such as headers and descriptions, thus increasing the dominance of morphological structures. Although a more thorough evaluation of parsing quality would have been desirable, we assume that the output is good enough for the present purposes.

B.3 Training Arguments

The GPT-2 model instances were fine-tuned using the `Transformers` library from Hugging Face. We used the pretrained `AutoTokenizer` with the padding token set to the end-of-sequence (EOS) token, with padding and truncation at a max length of 512. Each example in the data subset corresponds to one sentence, meaning that truncation is applied on the sentence level. The sentences have an average length of 40.08 tokens (with the caveat that many examples may consist of single titles or headers, which contribute to lowering this average), with 23,106 sentences (0.09%) exceeding the 512 tokens limit. We trained for 1 epoch using the `Trainer` API with an effective batch size of 32 (16 batches per device with gradient accumulation steps of 2), and the `AdamW` optimizer with a learning rate of 2e-5, a weight decay of 0.01 and half-precision (fp16) to speed up training.

[22] https://stanfordnlp.github.io/stanza/performance.html

Stage	Structure	Talbanken/LinES		OSCAR	
2	SVO	2,581	13,65%	4,208,314	11,66%
	PLU	479	2,53%	1,275,606	3,53%
	TNS	9,698	51,28%	18,871,751	52,29%
3	ATT	2,268	11,99%	4,816,125	13,34%
4	INV	3,258	17,23%	5,888,655	16,32%
	PR_a	226	1,20%	413,598	1,15%
	PR_b	4	0,02%	2,823	0,01%
5	NEG	304	1,61%	543,445	1,51%
	INQ	94	0,50%	70,230	0,19%

Table 5: Comparison of the distribution between PT structures in the Talbanken/LinES dataset and the training data (OSCAR after parsing 9%). Raw data is presented in the left columns, and the percentage of sentences annotated with each respective category in the right columns. A Chi-Square Test shows that the populations are significantly different. (X^2 (8, 17,874 = sample size) = 251.62, $p < 0.001$)

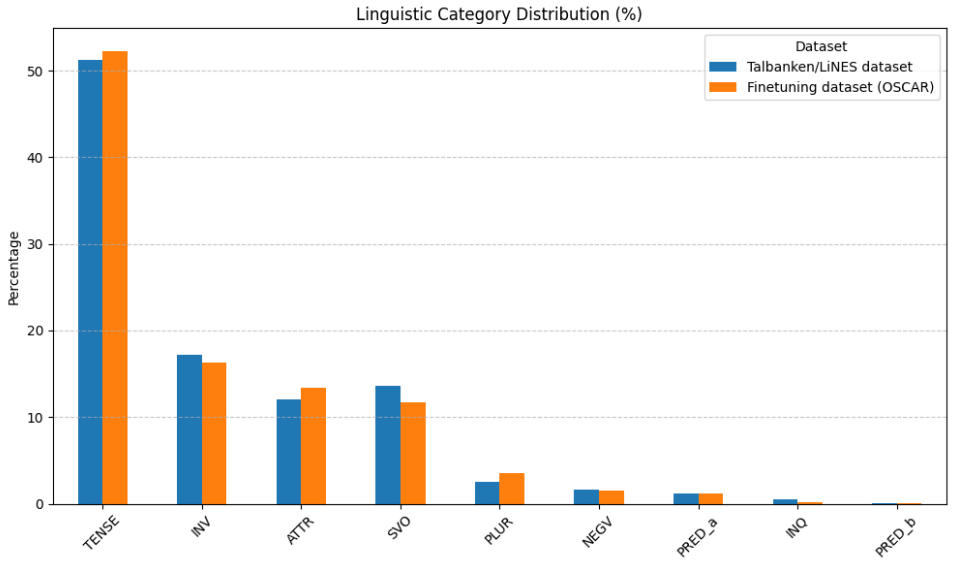


Figure 2: Distribution of linguistic categories between the evaluation dataset (SwePT) and the training data (subset from OSCAR) after parsing. The syntactic categories (SVO, INV, NEG, INQ) are consistently more frequent in the evaluation data, compared to in the training data. The opposite relationship is observed among the morphological structures (ATT, TNS, PLU, PR) which are independent from grammatical sentence structure.

C Acquisition thresholds

The acquisition threshold is determined per subset in relation to the number of minimal pairs (n) by finding the smallest number of correct guesses (k) possible to exceed chance level ($\pi = 50$) at the determined significance level ($\alpha = 0.05$). We used the reliability function representation, rewritten as $P(X \geq k) = \text{BinomSF}(k-1, n, \pi)$.

Structure	n	π=50%	π=60%	π=80%
SVO	2519	1302	1553	2049
PLU	479	258	306	398
TNS	2000	1038	1237	1630
ATT	2268	1174	1400	1847
PR_a	213	119	140	181
PR_b	27	19	21	26
INV	2581	1333	1590	2099
NEG	303	167	197	255
INQ	94	56	65	82

Table 6: Acquisition thresholds k per subset where $n =$ number of minimal pairs, and the remaining columns showing $k =$ the number of pairs needed to be correct to ensure acquisition above different thresholds π (with $\alpha = 0.05$).

D Additional Result Details

D.1 Implicational Scales

In our study, each checkpoint across all model instances are treated as individual learners in the implicational scales, thereby approximating a longitudinal PT study where the same learner is tested at multiple points throughout learning. Although treating each checkpoint independently may seem counter-productive considering that the purpose of this study is to examine developmental stages, implicational scaling aims to determine whether the observed acquisition order is implicational, that is, whether structures from lower developmental stages are *required* for the learner to acquire more complex structures. As such, the scales are agnostic to the specific point in time at which a learner is evaluated.

The columns represent the PT structures (9 in total), the rows represent the number of checkpoints sharing that particular acquisition order, and the cell values are coded as '+' (structure acquired) or '-' (structure not acquired). The columns (PT structures) are ordered according to their "overall rank order", i.e. from the most empirically difficult (the structure that has been acquired by the least learners) to the easiest (the structure that has been acquired by the most learners)[23]. The rows (the learners) are ordered from more advanced to less advanced in terms of how many PT structures they have acquired.

A coefficient or index of reproducibility (IR) is often used to give an indication of the scalability of the implicational scales, with a smaller IR expressing that a high number of entries deviate from the pattern, and a larger IR indicating a more consistent scale (1.0 reflecting a perfect scale). However, since multiple structures can be acquired during the same stage (for example, PT does not predict whether SVO is acquired before plural or tense since they are all processable at the same developmental stage), IR should be interpreted with caution. According to standard practice when using the PT framework, IR must be interpreted with consideration to the fact that not all structures of each stage must be emerged in order to consider that stage as acquired. However, this must be considered less relevant in the present study due to the large amount of data points in the evaluation dataset and the large number of checkpoints tested.

[23] In some PT studies (e.g., (Spinner, 2013)) the overall rank order is based on the order predicted by PT instead, where structures that belong to the same developmental stage are grouped together in the same category. However, due to the experimental nature of our study where the PT structures cannot be assumed but only hypothesized to fall within the same developmental stages, we use the standard approach previously described in favor of a more fine-grained analysis.

n	PR_b	PR_a	NEG	PLU	ATT	INV	INQ	SVO	TNS
4	-	+	+	+	+	+	+	+	+
235	-	-	+	+	+	+	+	+	+
21	-	+	-	+	+	+	+	+	+
3	-	+	+	-	+	+	+	+	+
99	-	-	-	+	+	+	+	+	+
96	-	-	+	-	+	+	+	+	+
11	-	+	-	+	+	-	+	+	+
15	-	-	-	-	+	+	+	+	+
26	-	-	-	+	-	+	+	+	+
15	-	-	-	+	+	-	+	+	+
4	-	-	+	-	-	+	+	+	+
1	-	-	+	-	+	+	-	+	+
49	-	-	-	-	-	+	+	+	+
6	-	-	-	-	+	+	-	+	+
30	-	-	-	-	-	-	+	+	+
5	-	-	-	-	-	+	-	+	+

(a) Acq. threshold: 50%. IR = 0.393

n	PR_a	PR_b	PLU	NEG	INV	ATT	INQ	TNS	SVO
30	-	-	+	+	+	+	+	+	+
194	-	-	-	+	+	+	+	+	+
3	-	-	+	-	+	+	+	+	+
81	-	-	-	-	+	+	+	+	+
32	-	-	-	+	-	+	+	+	+
17	-	-	-	+	+	+	-	+	+
58	-	-	-	-	-	+	+	+	+
12	-	-	-	-	+	+	-	+	+
3	-	-	-	+	-	+	-	+	+
1	-	-	-	+	+	-	-	+	+
131	-	-	-	-	-	-	+	+	+
19	-	-	-	-	-	+	-	+	+
5	-	-	-	+	-	-	-	+	+
28	-	-	-	-	-	-	-	+	+
2	-	-	-	+	-	-	-	-	+
4	-	-	-	-	-	-	-	-	+

(b) Acq. threshold: 60%. IR = 0.292

n	PLU	PR_a	PR_b	INV	NEG	INQ	ATT	TNS	SVO
3	-	-	-	-	+	+	-	+	+
12	-	-	-	-	-	+	-	+	+
2	-	-	-	-	+	-	-	+	+
254	-	-	-	-	-	-	-	+	+
41	-	-	-	-	-	-	+	-	+
195	-	-	-	-	-	-	-	-	+
7	-	-	-	-	+	-	-	-	-
106	-	-	-	-	-	-	-	-	-

(c) Acq. threshold: 80%. IR = 0.697

Table 7: Implicational scales across all checkpoints from the four models evaluated on SwePT, based on acquisition times calculated at three different acquisition thresholds. A '+' mark indicates that that specific structure is acquired. A '-' mark indicates that the structure is not acquired. The scales are collapsed, meaning that checkpoints with identical acquisition order are counted together in the same row. n denotes this number of checkpoints. The structures are ordered from left to right, with the structure that is learned at the most checkpoints to the left, and the structure learned by the least checkpoints to the right. IR = Index of Reproducibility

E Acquisition times

Model	SVO	PLU	TNS	ATT	PR_a	PR_b	INV	NEG	INQ
GPT-mixed	1	100	1	10	-	-	1	100	10
GPT-mixed_2	1	100	1	10	-	-	1	100	10
GPT-order	1	100	1	100	100	-	1	154	10
GPT-reverse	1	100	1	10	53	-	1	1	1

(a) Acquisition threshold set at 50% accuracy.

Model	SVO	PLU	TNS	ATT	PR_a	PR_b	INV	NEG	INQ
GPT-mixed	1	128	10	10	-	-	100	100	100
GPT-mixed_2	1	126	10	100	-	-	100	100	100
GPT-order	1	142	1	100	-	-	109	154	100
GPT-reverse	1	-	10	10	-	-	10	1	100

(b) Acquisition threshold set at 60% accuracy.

Model	SVO	PLU	TNS	ATT	PR_a	PR_b	INV	NEG	INQ
GPT-mixed	100	-	100	-	-	-	-	-	100
GPT-mixed_2	100	-	100	-	-	-	-	-	-
GPT-order	100	-	117	100	-	-	-	154	156
GPT-reverse	100	-	100	56	-	-	-	2	109

(c) Acquisition threshold set at 80% accuracy.

Table 8: The time of acquisition per structure and model, calculated at thresholds of 50%, 60% and 80% accuracy (see Table 6 for the calculation of each structure-specific threshold). The numbers indicate the checkpoints, which are saved at intervals of 100 time steps. A dash indicates that that structure has never reached the respective threshold (not acquired).

Threshold	Mean correlation	p-value
50%	0.8373	0.0000
60%	0.7214	0.0000
80%	0.8833	0.0110

Table 9: Results from the rank correlation permutation test across all four models. The high mean correlation scores indicate that the models acquire the grammatical structures in a consistent order. The low p-values indicate high significance of this correlation. The high correlation but low significance for the 80% accuracy threshold is explained by the lower amount of data points in that group.

E.1 Acquisition trajectories

E.2 Model confidence on acceptability scores

Figures 4 and 5 plot the absolute differences between the log-likelihood scores assigned to the grammatical vs. the ungrammatical sentence of each minimal pair from the AJT, i.e. the models' "confidence" in their classifications across checkpoints. While following similar envelopes as the learning curves, the confidence curves are smoother, indicating that confidence in grammatical distinctions improves more gradually and is less sensitive to noise or outliers, than raw loss.

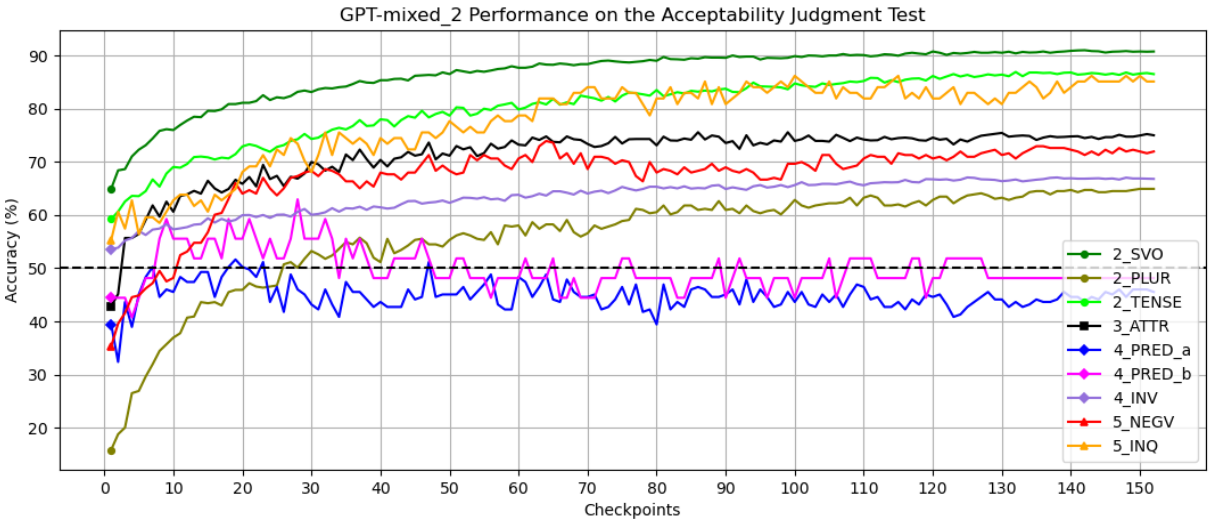
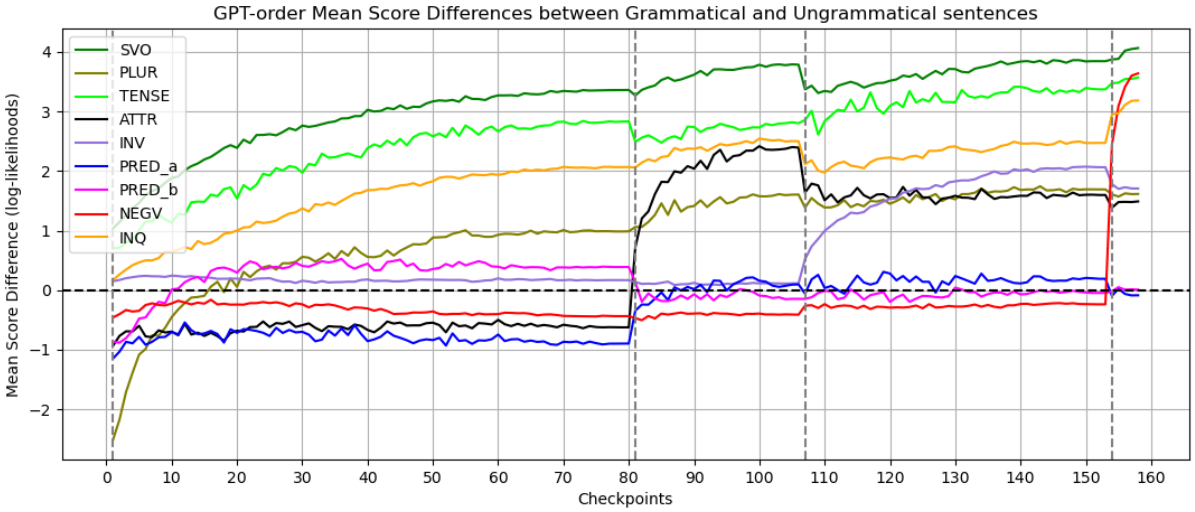
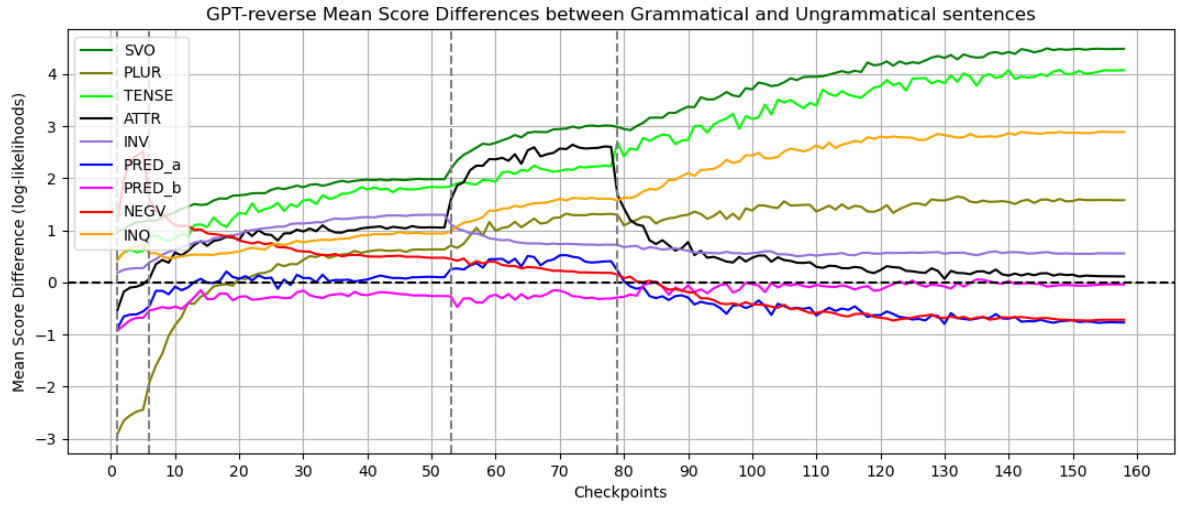


Figure 3: Results from the AJT on all linguistic structures across checkpoints for the model trained on all four data subsets concatenated and shuffled (seed=123). The acquisition trajectories follow a relatively parallel pattern, and the acquisition order deviates from that predicted by PT.

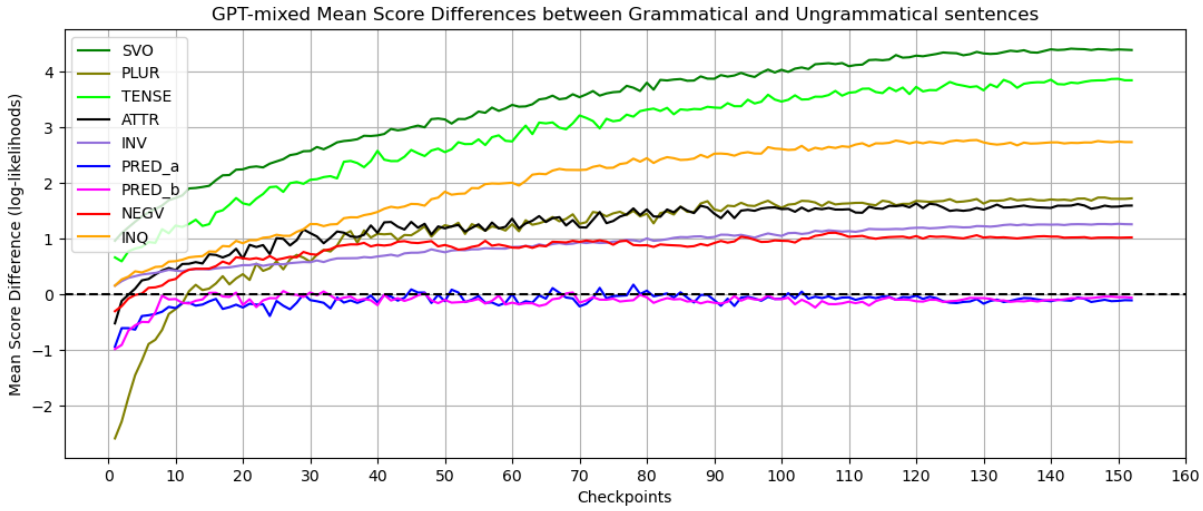


(a) Absolute differences for the GPT-order model.

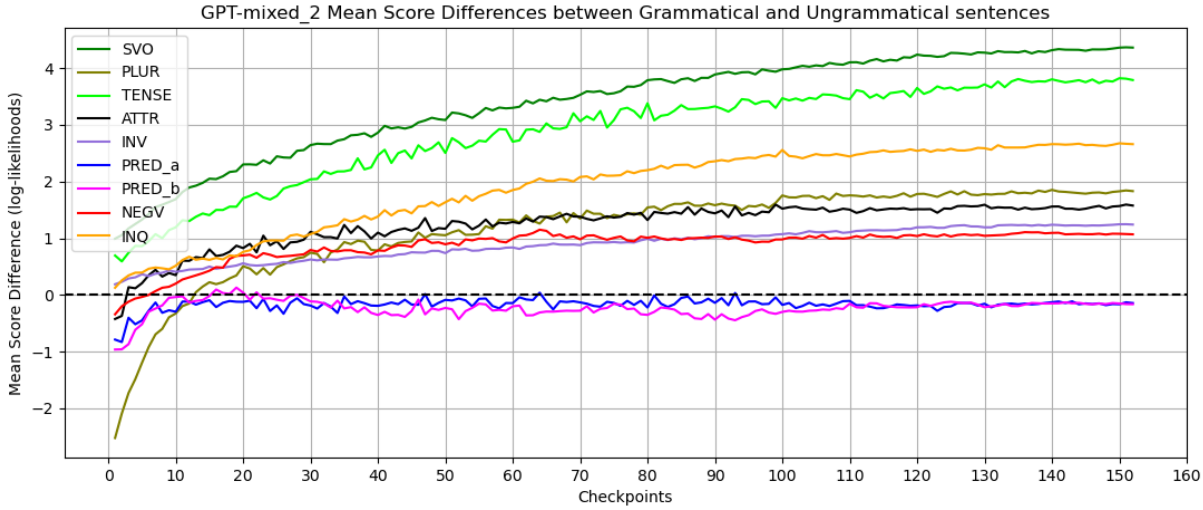


(b) Absolute differences for the GPT-reverse model.

Figure 4: Absolute differences between the log-likelihood scores assigned to the grammatical and ungrammatical sentences of the minimal pairs from the AJT for the curriculum models. The vertical dashed lines mark where training commences on data from a new Stage subset.



(a) Absolute differences for the GPT-mixed model.



(b) Absolute differences for the second GPT-mixed model.

Figure 5: Absolute differences between the log-likelihood scores assigned to the grammatical and ungrammatical sentences of the minimal pairs from the AJT for the mixed models.

On the Learnability of Syntax from Raw Speech with Autoregressive Predictive Coding

Shunsuke Kando, Yusuke Miyao
Department of Computer Science, The University of Tokyo,
{skando,yusuke}@is.s.u-tokyo.ac.jp

Abstract

Children are known to generalize syntactic knowledge at ages when their linguistic input is predominantly raw speech rather than text. This raises the question of whether syntactic generalization can emerge directly from acoustic input. We address this question using Autoregressive Predictive Coding (APC), a simple prediction-based self-supervised speech model. To approximate the input available to human learners while enabling controlled comparison, we train models on both child-directed speech and audiobook speech. We evaluate the models on a minimal-pair benchmark targeting elementary syntactic phenomena, designed to be acquisition-friendly. Our results show that APC partially generalizes word-order regularities when trained to predict near-future frames. However, the model fails to generalize agreement phenomena, suggesting that predictive learning from acoustic signals alone is insufficient. Furthermore, we observe distinct learning dynamics across word-order phenomena, suggesting that some improvements may be driven by shallow statistical regularities rather than genuine syntactic generalization.

1 Introduction

Children acquire a wide range of linguistic knowledge from speech input. Starting with sensitivity to prosody at birth (Mehler et al., 1988), they are reported to acquire typical vowel categories by 6 months (Kuhl et al., 1992), common nouns by 9 months (Bergelson and Swingley, 2012), and clausal units by 10 months (Hirsh-Pasek et al., 1987). Higher-level syntactic knowledge, such as agreement and word order, is typically acquired around the age of 3–4 (Kenney and Wolfe, 1972; Akhtar, 1999). Although such knowledge takes longer to develop than lower-level linguistic abilities, children's input remains predominantly raw speech rather than text throughout this period. A central question in developmental linguistics is how children can generalize a wide range of linguistic knowledge from noisy and limited speech input. While behavioral experiments provide insights into what children know about language, they do not directly reveal how such knowledge is acquired (Rowland et al., 2025). Moreover, such studies are often constrained by individual variability and limited scalability, making it difficult to obtain systematic and reproducible evidence.

In contrast to behavioral experiments, computational modeling offers a complementary approach for investigating the mechanisms of language acquisition in a controlled and scalable manner (Räsänen, 2026). While classical computational approaches have tested the language learning hypothesis primarily with artificial data or transcribed speech (Elman, 1990; Aslin et al., 1996; de Marcken, 1996), recent advances in neural network models enable simulation from realistic input, including raw speech (Dupoux, 2018; Räsänen, 2026). Previous research has examined phonetic/lexical learning (Lavechin et al., 2025; Khorrami et al., 2023), structural prosodic knowledge (De Seyssel et al., 2023), and word or syllable segmentation (Algayres et al., 2022; Pasad et al., 2024; Cho et al., 2025; Baade et al., 2025), using raw audio or audiovisual input. However, most work that examines higher-level linguistic knowledge involves optimization of multiple components (e.g., representation learning and language modeling (Lavechin et al., 2025)), making it difficult to isolate which components contribute to the observed performance.

In this paper, we examine the learnability of syntax with a predictive coding model. We employ Autoregressive Predictive Coding (APC; Chung et al., 2019), a simple yet effective self-supervised learning method. To evaluate syntactic knowledge, we use BabySLM (Lavechin et al., 2023), a minimal-pair benchmark designed to capture elementary syntactic phenomena. In the original pa-

Proceedings of the 1st Workshop on Computational Developmental Linguistics (CDL), pages 77–82

per, Lavechin et al. (2023) trains both representation learning and language modeling, where the latter models sequential patterns in the learned representations. In contrast, our approach isolates representation learning by removing language modeling over discrete symbolic units, allowing us to directly assess whether predictive coding alone can induce syntactic generalization. We trained models on both audiobook and child-directed speech audio.

Our results show that certain word-order phenomena can be learned with APC, although the specific patterns that are acquired depend on the training data. In contrast, agreement phenomena remain at chance level, suggesting that predictive coding alone is insufficient for capturing more complex syntactic dependencies. Beyond final performance, we also analyze the learning trajectory of syntactic phenomena over the course of training. We observe that the accuracy on Adj–noun order rises above chance at the beginning of training (far before 1 epoch), which might indicate the model's reliance on shallow local regularities rather than genuine syntactic generalization. Furthermore, different word-order phenomena exhibit distinct learning dynamics, indicating that not all apparent improvements reflect the same type of generalization.

The experimental codebase is made publicly available[1].

2 Autoregressive Predictive Coding

Autoregressive Predictive Coding (APC; Chung et al., 2019) is an unsupervised representation learning method based on predictive coding. Given a sequence of log Mel spectrograms $\boldsymbol{x} = (x_1, x_2, \ldots, x_T)$, the model is optimized to predict a frame n step ahead of the current one. Prediction vectors $\boldsymbol{y} = (y_1, y_2, \ldots, y_T)$ are produced by a recurrent neural network (RNN). The objective is defined as the L1 loss between the predicted and target frames:

$$\mathcal{L} = \sum_{i=1}^{T-n} |x_{i+n} - y_i|,$$

where n denotes the number of frames ahead to predict. The hyperparameter n encourages the model to capture global structures beyond immediate frame-level continuity.

Despite its simplicity and unsupervised nature, APC has been shown to be effective across a wide range of speech tasks, including phone classification and discrimination (Blandón and Räsänen, 2020), speaker verification (Chung et al., 2019), and automatic speech recognition (Yang et al., 2022).

A major variant of APC is Contrastive Predictive Coding (CPC) (van den Oord et al., 2019), which differs in two key aspects: (1) it employs a convolutional neural network for feature extraction instead of Mel spectrograms, and (2) it is trained with a contrastive loss instead of an L1 loss. In this work, we focus on APC due to its architectural simplicity and interpretability as a predictive model, which makes it suitable for analyzing the learnability of syntactic structure. We note that CPC has also been widely studied in the context of language acquisition modeling, and extending our analysis to CPC remains an important direction for future work.

3 Experimental Setup

To examine whether developmentally plausible speech audio can induce syntactic knowledge, we trained models on a child-directed speech (CDS) dataset. In addition, we train separate models on audiobook speech as a contrasting condition to assess the role of input characteristics. Syntactic knowledge is evaluated using a minimal-pair benchmark.

3.1 Dataset

To construct the CDS dataset, we extract English subsets[2] from CHILDES database (Macwhinney, 2000). We trim the spoken part of the audio using time alignments provided in the CHAT transcriptions. Since our focus is on child-directed speech audio, we exclude utterances produced by children or speakers with unknown roles. The resulting CDS dataset has a total duration of 995 hours. As a contrasting condition, we use LibriSpeech (Panayotov et al., 2015) as an audiobook dataset, with a comparable total duration of 960 hours. The total number of utterances is 2M for CDS and 281K for audiobook, indicating that CDS consists of shorter utterances on average. The dataset is split into training set and validation set at the ratio of 99:1.

3.2 Model Setup

We use an APC model consisting of three unidirectional LSTM layers with 512 hidden units each, followed by a linear layer for frame prediction. We observe that the number of training steps required

[1] https://github.com/gifdog97/babyslm_apc

[2] Eng-AAE, Eng-NA, and Eng-UK.

Table 1: Example pairs of syntactic tasks in BabySLM.

Phenomenon	Pair example
Adjective–noun order	✓ The good mom. ✗ The mom good.
Noun–verb order	✓ The dragon says. ✗ The says dragon.
Anaphor–gender agreement	✓ The dad cuts himself. ✗ The dad cuts herself.
Anaphor–number agreement	✓ The boys told themselves. ✗ The boys told himself.
Determiner–noun agreement	✓ Each good sister. ✗ Many good sister.
Noun–verb agreement	✓ The prince needs the princess. ✗ The prince need the princess.

for convergence differs between CDS and audiobook data, likely due to differences in the number and length of utterances. Hence, we train for 5 epochs on CDS and 25 epochs on the audiobook dataset. We used an AdamW optimizer with a batch size of 256 and an initial learning rate of 10^{-3}. We varied the step size n from 1 to 16.

3.3 Evaluation

To evaluate syntactic knowledge, we use BabySLM (Lavechin et al., 2023), a minimal-pair benchmark targeting elementary syntactic phenomena (Table 1). In this setup, each pair consists of a grammatical and an ungrammatical sentence. The model is evaluated based on whether it assigns a higher score to the grammatical audio. In the original BabySLM setup, language models are trained on top of learned speech representations, and scores are computed using negative log-likelihood. In contrast, we directly use the prediction error of APC as a scoring function. Specifically, the score of an input audio sequence $\boldsymbol{x}$ is defined as:

$$score(\boldsymbol{x}) = -\frac{1}{T-n}\sum_{i=1}^{T-n}|x_{i+n} - y_i|,$$

where higher scores indicate better predictions.

4 Results

4.1 The effect of dataset and n

Table 2 shows the average accuracy across syntactic phenomena. Overall, most accuracies remain close to chance level, regardless of the dataset and the prediction step n. This confirms the inherent difficulty of the syntactic tasks, which has also been

Table 2: Overall accuracy on BabySLM syntactic test. Chance rate is 0.5.

n	CDS		AudioBook	
	dev	test	dev	test
1	0.52	0.505	0.51	0.502
2	0.533	0.513	0.527	0.51
3	0.528	0.501	0.535	0.502
4	0.518	0.5	0.537	0.509
5	0.491	0.479	0.548	0.504
6	0.498	0.477	0.523	0.495
7	0.494	0.472	0.507	0.491
8	0.464	0.469	0.515	0.504
9	0.467	0.464	0.494	0.494
10	0.468	0.461	0.495	0.502
11	0.459	0.46	0.494	0.51
12	0.445	0.444	0.488	0.49
13	0.456	0.458	0.472	0.494
14	0.45	0.457	0.462	0.482
15	0.469	0.475	0.469	0.485
16	0.459	0.455	0.497	0.493

observed in prior work on speech-based language models (Lavechin et al., 2023). Figure 1 shows accuracy broken down by syntactic phenomenon on the development set. These results show that certain word-order phenomena can be partially learned with APC. In particular, Adj–noun order is captured when models are trained on child-directed speech with near-feature prediction ($2 \leq n \leq 7$); Noun–verb order is captured under a narrower range ($3 \leq n \leq 5$) when trained on audiobook data. In contrast, models trained with $n = 1$ fail to capture word-order regularities, indicating that immediate frame-level prediction is insufficient and that integrating information over a slightly longer temporal context is necessary.

On the other hand, all agreement phenomena remain at near-chance levels across all values of n and training data. This suggests that predictive learning from local acoustic signals alone is insufficient to capture agreement, which involves dependencies between more abstract linguistic units. This may indicate the potential importance of modeling over discrete or symbolic units, as in language models.

Interestingly, performance on Noun–verb order falls significantly below chance at larger values of n, suggesting a preference for the opposite order.

4.2 Analysis of Learning Trajectory

To better understand the nature of syntactic learning of APC, we analyze the learning trajectory of each syntactic phenomenon. We track the accuracy for the $n = 3$ setting throughout training, evaluating every 100 steps from 0 to 6,300. Figure 2

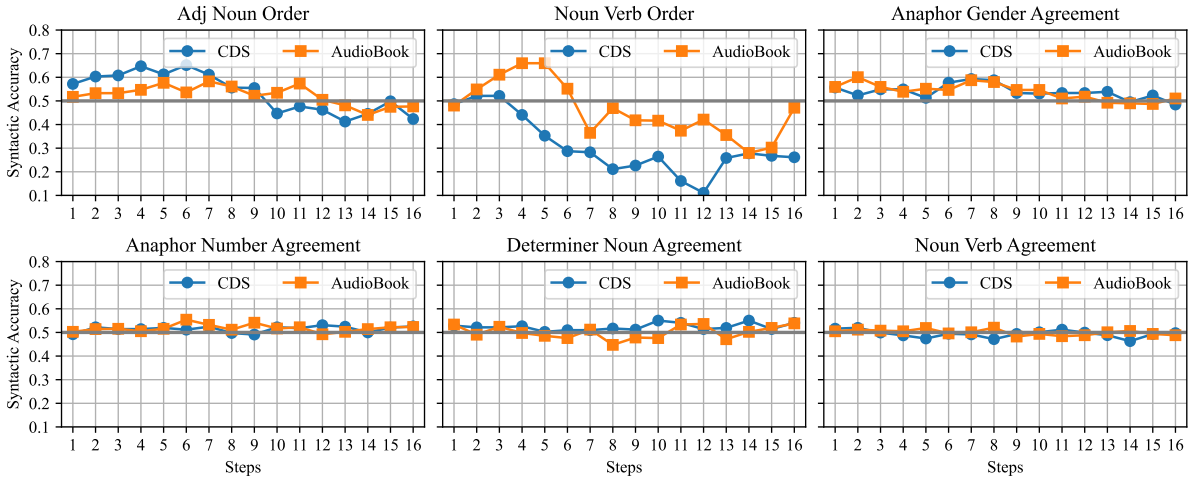


Figure 1: Accuracy of each phenomenon. X-axis represents n, the number of frame steps ahead to predict.

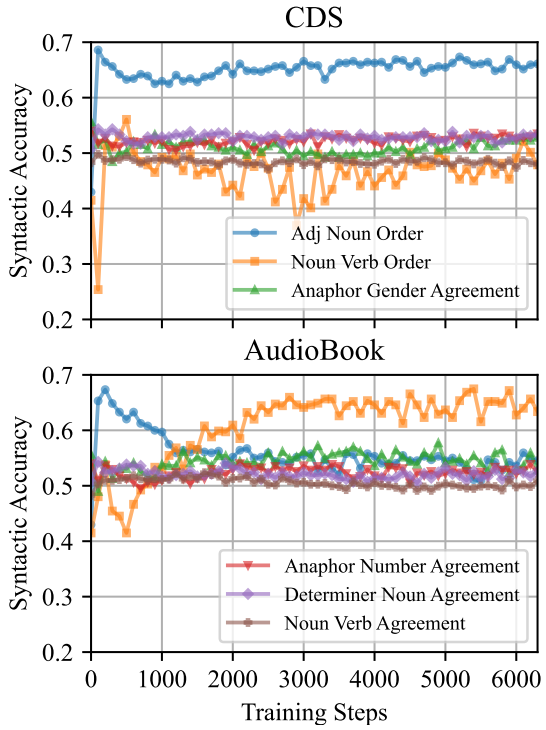


Figure 2: The trajectory of accuracy with $n = 3$.

shows the result. We observe that the accuracy of agreement phenomena remains at near-chance levels throughout training, further supporting the difficulty of learning it with predictive coding. In contrast, word-order phenomena exhibit distinct learning dynamics. For both datasets, the accuracy of Adj–noun order increases rapidly at the first 100 steps, well before completing a single epoch[3]. One possible explanation for this behavior is that the model relies on shallow local regularities (e.g., local co-occurrence patterns) rather than genuine syntactic knowledge. Since APC is optimized for local acoustic prediction, it may exploit short-range patterns that correlate with the correct answer. Such cues may not reflect the word-order structure that the benchmark is intended to probe. Further investigation is required to clarify whether the model genuinely acquires syntactic generalization.

We also observe the swap of accuracy between Adj–noun order and Noun–verb order when models are trained on audiobook data. This contrast suggests that the two phenomena may rely on qualitatively different cues. The rapid rise and subsequent degradation in Adj–noun order accuracy are consistent with the model relying on shallow local regularities that do not generalize beyond early training. In contrast, the gradual improvement observed in Noun–verb order may reflect the acquisition of more robust and generalizable patterns. In this sense, Noun–verb order might provide a more reliable indicator of genuine generalization than Adj–noun order.

5 Conclusion

In this paper, we investigated the learnability of syntax from raw speech using Autoregressive Predictive Coding (APC). While APC captures certain word-order phenomena, it fails to generalize agreement, suggesting that predictive coding over local acoustic signals is insufficient for modeling complex syntactic dependencies. Analysis of learning trajectories further reveals that rapid improvements may stem from shallow statistical regularities rather than genuine generalization. These results highlight the need for additional mechanisms, such

[3]1 epoch equals 7,600 steps for CDS and 1,000 steps for audiobook.

as representations over more abstract or symbolic units, for acquiring syntactic knowledge.

Limitations

First, we focused on a single predictive coding model (APC), whilce CPC (van den Oord et al., 2019) has also been widely studied in speech representation learning or language acquisition modeling literature. Future work should examine CPC and related models to determine whether our findings generalize beyond APC. Second, we used only BabySLM benchmark for evaluation. To better assess fine-grained syntactic knowledge of the models, future work should complement this benchmark with additional evaluation methods, such as probing (He et al., 2025) or canonical correlation analysis (Pasad et al., 2024). Third, the cause of the distinct learning trajectories observed for word-order phenomena remains unclear. In particular, our hypothesis that Adj–noun order may be solvable without syntactic generalization requires more careful verification. Finally, the training dataset is not fully developmentally plausible. Children are reported to be exposed not only to CDS, but also to overheard speech (Thompson, 2018) and media audio (Gowenlock et al., 2024), both of which constitute a non-negligible portion of their linguistic input. In addition, the audio quality of CHILDES recordings is not always ideal, particularly for older corpora. To better facilitate the reverse engineering of language acquisition, further work is needed to construct more ecologically valid models of children's language inputs.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 26KJ0792 and JST ACT-X Grant Number JPMJAX24C9.

References

Nameera Akhtar. 1999. Acquiring basic word order: Evidence for data-driven learning of syntactic structure. *Journal of child language*, 26:339–56.

Robin Algayres, Tristan Ricoul, Julien Karadayi, Hugo Laurençon, Salah Zaiem, Abdelrahman Mohamed, Benoît Sagot, and Emmanuel Dupoux. 2022. DP-Parse: Finding Word Boundaries from Raw Speech with an Instance Lexicon. *Transactions of the Association for Computational Linguistics*, 10:1051–1065.

Richard N. Aslin, Julide Z. Woodward, Nicholas P. LaMendola, and Thomas G. Bever. 1996. Models of word segmentation in fluent maternal speech to infants. In *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*, pages 117–134. Lawrence Erlbaum Associates, Inc.

Alan Baade, Puyuan Peng, and David Harwath. 2025. SyllableLM: Learning Coarse Semantic Units for Speech Language Models. In *ICLR 2025*. OpenReview.net.

Elika Bergelson and Daniel Swingley. 2012. At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109(9):3253–3258.

María Andrea Cruz Blandón and Okko Räsänen. 2020. Analysis of Predictive Coding Models for Phonemic Representation Learning in Small Datasets. In *workshop on Self-supervision in Audio and Speech at ICML 2020*.

Cheol Jun Cho, Nicholas Lee, Akshat Gupta, Dhruv Agarwal, Ethan Chen, Alan W. Black, and Gopala K. Anumanchipalli. 2025. Sylber: Syllabic Embedding Representation of Speech from Raw Audio. In *ICLR 2025*. OpenReview.net.

Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. 2019. An Unsupervised Autoregressive Model for Speech Representation Learning. In *INTERSPEECH 2019*, pages 146–150.

Carl de Marcken. 1996. *Unsupervised Language Acquisition*. phdthesis, MIT.

Maureen De Seyssel, Marvin Lavechin, Hadrien Titeux, Arthur Thomas, Gwendal Virlet, Andrea Santos Revilla, Guillaume Wisniewski, Bogdan Ludusan, and Emmanuel Dupoux. 2023. ProsAudit, a prosodic benchmark for self-supervised speech models. In *INTERSPEECH 2023*, pages 2963–2967. ISCA.

Emmanuel Dupoux. 2018. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59.

Jeffrey L. Elman. 1990. Finding Structure in Time. *Cognitive Science*, 14(2):179–211.

Anna Elizabeth Gowenlock, Courtenay Norbury, and Jennifer M. Rodd. 2024. Exposure to language in video and its impact on linguistic development in children aged 3–11: A scoping review. *Journal of Cognition*.

Linyang He, Qiaolin Wang, Xilin Jiang, and Nima Mesgarani. 2025. Layer-wise Minimal Pair Probing Reveals Contextual Grammatical-Conceptual Hierarchy in Speech Representations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35338–35353. Association for Computational Linguistics.

Kathy Hirsh-Pasek, Deborah G. Kemler Nelson, Peter W. Jusczyk, Kimberly Wright Cassidy, Benjamin Druss, and Lori Kennedy. 1987. Clauses are perceptual units for young infants. *Cognition*, 26(3):269–286.

Terrence J. Kenney and Jean Wolfe. 1972. The acquisition of agreement in English. *Journal of Verbal Learning and Verbal Behavior*, 11(6):698–705.

Khazar Khorrami, María Andrea Cruz Blandón, and Okko Räsänen. 2023. Computational Insights to Acquisition of Phonemes, Words, and Word Meanings in Early Language: Sequential or Parallel Acquisition? *Proceedings of the Annual Meeting of the Cognitive Science Society*, 45(45).

P. K. Kuhl, K. A. Williams, F. Lacerda, K. N. Stevens, and B. Lindblom. 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255(5044):606–608.

Marvin Lavechin, Maureen de Seyssel, Hadrien Titeux, Guillaume Wisniewski, Hervé Bredin, Alejandrina Cristia, and Emmanuel Dupoux. 2025. Simulating Early Phonetic and Word Learning Without Linguistic Categories. *Developmental Science*, 28(2):e13606.

Marvin Lavechin, Yaya Sy, Hadrien Titeux, María Andrea Cruz Blandón, Okko Räsänen, Hervé Bredin, Emmanuel Dupoux, and Alejandrina Cristia. 2023. BabySLM: Language-acquisition-friendly benchmark of self-supervised spoken language models. In *INTERSPEECH 2023*, pages 4588–4592. ISCA.

Brian Macwhinney. 2000. The CHILDES Project: Tools for Analyzing Talk (third edition): Volume I: Transcription format and programs, Volume II: The database. *Computational Linguistics*, 26:657–657.

Jacques Mehler, Peter Jusczyk, Ghislaine Lambertz, Nilofar Halsted, Josiane Bertoncini, and Claudine Amiel-Tison. 1988. A precursor of language acquisition in young infants. *Cognition*, 29(2):143–178.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu. 2024. What Do Self-Supervised Speech Models Know About Words? *Transactions of the Association for Computational Linguistics*, 12:372–391.

Caroline F. Rowland, Gert Westermann, Anna L. Theakston, Julian M. Pine, Padraic Monaghan, and Elena V. M. Lieven. 2025. Constructing language: A framework for explaining acquisition. *Trends in Cognitive Sciences*.

Okko Räsänen. 2026. Computational modeling of early language learning from acoustic speech and audiovisual input without linguistic priors. *Preprint*, arXiv:2603.08359.

Abbie Thompson. 2018. Who's Talking to Whom and Does It Matter? The Impact of Multiple Speakers, Overheard Speech, and Child-Directed Speech on Infants' Language Development.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation Learning with Contrastive Predictive Coding. *Preprint*, arXiv:1807.03748.

Gene-Ping Yang, Sung-Lin Yeh, Yu-An Chung, James Glass, and Hao Tang. 2022. Autoregressive Predictive Coding: A Comprehensive Study. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1380–1390.

Modeling Writing Development as Coordinated Change Across Linguistic and Semantic Dimensions

Michelle Banawan, Andrew Potter,
Tracy Arner, and **Danielle S. McNamara**

Learning Engineering Institute
Arizona State University
Tempe, AZ, USA
{mbanawan, ahpotter, tarner, dsmcnama}@asu.edu

Abstract

Writing development is often assessed through aggregate improvements in surface-level features, yet less attention has been given to how multiple linguistic dimensions evolve jointly over time. We model writing development as a multidimensional system shaped by stable individual variation and instructional progression across staged assignments, using interpretable linguistic features from the Writing Analytics Toolkit (WAT) and transformer-based sentence embeddings.

Variance partitioning reveals substantial between-student stability alongside stage-dependent change. Mixed-effects models identify non-uniform developmental trajectories: academic focus, information density, and conventional language increase, whereas development of ideas and lexical variety decline, indicating tradeoffs across competing dimensions. Cross-lagged analyses further show dynamic dependencies between dimensions, suggesting coordinated change rather than independent progression.

Embedding-based analyses capture stage-dependent shifts in semantic representation, with larger changes in earlier stages and greater stability in later stages. Although assignment structure contributes to observed variation, stable learner-level variation and cross-stage dependencies extend across instructional tasks.

Together, these findings characterize writing development as structured change in a multidimensional representational system, highlighting the need for computational models that capture stable variation, non-monotonic trajectories, and interactions among linguistic components.

1 Introduction

Understanding how writing develops among college writers under instruction is central to both educational research and computational models of language learning. Prior computational and applied linguistics research emphasizes that writing reflects interactions between individual linguistic repertoires and learning environments (Crossley, 2020; Crossley and Kim, 2022). Computational studies further show that writing exhibits stable individual variation across linguistic and representational spaces, although different representational frameworks capture distinct aspects of writing, including stylistic, functional, and semantic organization (Zhu and Jurgens, 2021). However, most computational approaches evaluate language ability cross-sectionally, with limited attention to how linguistic abilities evolve over time in structured instructional contexts (Crossley, 2020). Writing development is inherently multidimensional, involving coordinated changes across interacting linguistic dimensions that may evolve at different rates (Crossley and Kim, 2022; Goulart and Wood, 2021).

In academic settings, writing instruction typically progresses through sequenced assignments of increasing rhetorical complexity, moving from structured comparison to descriptive elaboration and research-based argumentation. These transitions provide a natural framework for examining development across instructional milestones. However, existing work often focuses on overall improvement or isolated features, rather than modeling the organization and interaction of multiple linguistic dimensions within a staged curriculum.

This study models writing development using a repeated-measures dataset of student essays collected across five instructional stages in a first-year university writing course, replicated across three sections. Each essay is represented using nine interpretable functional indices (component scores) derived from automated linguistic analysis, including Sophisticated Wording, Development of Ideas, Word Variety, Information Density, Academic Focus, and Conversational Writing Style. These in-

Proceedings of the 1st Workshop on Computational Developmental Linguistics (CDL), pages 83–91

dices capture higher-order functional properties of writing rather than surface-level counts.

We address three research questions: (1) How is variance in writing dimensions partitioned across students and course sections? (2) Do these dimensions exhibit systematic change across instructional stages? (3) Are dimensions dynamically coupled, such that prior values of one dimension predict subsequent change in another?

To answer these questions, we integrate variance partitioning, growth modeling, and cross-lagged analyses to characterize writing development as a multidimensional system shaped by stable individual variation and stage-dependent reorganization.

From a computational perspective, we model writing development as representational change over time, examining how texts evolve across interpretable linguistic features and neural representations rather than treating writing ability as a single outcome. We combine WAT-derived linguistic features with transformer-based sentence embeddings to link interpretable linguistic analysis with representation learning. Linguistic features capture variation and interactions across dimensions of writing, whereas embeddings reflect stage-dependent changes in semantic representation.

This work makes three contributions. First, it models writing development as coordinated change across interacting linguistic dimensions within a structured instructional setting. Second, it provides empirical evidence that developmental trajectories are non-uniform and involve tradeoffs between competing dimensions. Third, it links interpretable linguistic features with neural representations, showing how each captures distinct aspects of development.

2 Feature Spaces for Modeling Writing Development

2.1 Functional Linguistic Dimensions (WAT Components)

Functional linguistic dimensions are derived from the Writing Analytics Toolkit (WAT), which organizes large sets of linguistic features into a smaller set of interpretable dimensions using principal component analysis (PCA) (Potter et al., 2026). PCA identifies patterns of covariance among features and reduces them into components representing broader functional properties of writing.

The models are trained on a corpus of academic writing, enabling identification of stable patterns across genres (McNamara et al., 2026). The resulting components capture constructs such as academic language use, cohesion, elaboration, and lexical variety.

In this study, component scores are used as interpretable indices of writing. Because each dimension reflects coordinated variation among multiple features, changes across instructional stages are interpreted as shifts in underlying linguistic subsystems rather than isolated feature-level variation.

2.2 Neural Embedding Representation of Student Writing

Neural embeddings represent text as dense vectors that encode semantic and contextual relationships learned from large corpora (Devlin et al., 2019). We use Sentence-BERT (SBERT), which produces fixed-length sentence embeddings optimized for semantic similarity, enabling efficient comparison using cosine similarity (Reimers and Gurevych, 2019).

Embedding-based representations have been applied to student writing to capture semantic similarity and discourse structure, complementing feature-based approaches that emphasize surface-level variation (Fiacco et al., 2022). However, transformer-based semantic embeddings are known to entangle topical and stylistic information, and are therefore not optimized for isolating stable authorial style independently of content (Wegmann et al., 2022).

In this study, embeddings provide a complementary representation of student writing alongside WAT-derived features. While WAT components capture interpretable variation across linguistic dimensions, embeddings encode the overall semantic structure of texts and are used to analyze trajectories in semantic representation across instructional stages, capturing how assignment structure shapes writing over time.

3 Methods

3.1 Data Construction and Longitudinal Design

The dataset was constructed from institutional Canvas exports containing submission-level metadata for student writing assignments, including assignment identifiers, timestamps, student identifiers, and course section identifiers (McNamara et al., 2022). We restrict the analysis to three sections of a first-year university writing course (English 101) that implemented an identical staged curriculum.

After filtering to include final draft submissions from students who completed the course and consolidating records, the final dataset comprises 308 student essays produced by 62 students across five instructional stages and three course sections, with essays serving as the unit of analysis.

The curriculum consists of five sequenced writing assignments of increasing rhetorical complexity: Compare/Contrast, Illustration, Descriptive, Persuasive Research, and a discussion-based synthesis assignment. These assignments reflect standard genres in introductory composition courses, progressing from structured comparison to descriptive elaboration and research-based argumentation, and provide a structured framework for examining development across instructional stages.

Assignments are mapped to a stage index ($stage\_index \in \{1, 2, 3, 4, 5\}$) based on their curricular order. Because each stage is implemented across all sections, the dataset forms a partially replicated longitudinal design.

Stage 1 corresponds to the Compare/Contrast assignment, Stage 2 to Illustration, Stage 3 to Descriptive writing, Stage 4 to Persuasive Research, and Stage 5 to the synthesis assignment.

Each observation corresponds to a single student submission at a given stage and includes assignment metadata, section identifiers, and the stage index used in subsequent analyses.

3.2 Linguistic Feature Extraction

Linguistic features were extracted from each essay using WAT. PCA-derived component scores were used as functional indices of writing, with each score representing a text's position along an underlying linguistic dimension rather than the frequency of a single feature.

The nine dimensions analyzed are: Sophisticated Wording, Sentence Cohesion, Word Variety, Conversational Writing Style, Academic Focus, Conventional Language, Information Density, Word Concreteness, and Development of Ideas. These dimensions capture higher-order properties of writing, including lexical sophistication, stylistic orientation, informational density, and discourse elaboration.

Because each dimension reflects coordinated variation among multiple features, changes across instructional stages are interpreted as shifts in underlying linguistic subsystems rather than isolated feature-level variation.

3.3 Embedding-based Modeling

To model writing at the representational level, essays were encoded using Sentence-BERT (SBERT) with the `all-mpnet-base-v2` pretrained model (Reimers and Gurevych, 2019), a transformer-based model optimized for semantic similarity. Because essay-length inputs exceeded the model's maximum token length, essays were segmented into smaller text units prior to encoding. Embeddings were generated for each segment and aggregated using mean pooling to obtain a single essay-level semantic representation. The resulting embeddings capture contextual and semantic relationships across the full document in a high-dimensional space (Devlin et al., 2019).

Representational change across instructional stages was quantified by computing cosine similarity between embeddings of consecutive essays written by the same student:

$$sim(v_i, v_j) = \frac{v_i \cdot v_j}{\|v_i\| \, \|v_j\|} \tag{1}$$

Representational change was operationalized as:

$$drift = 1 - sim(v_i, v_j) \tag{2}$$

Higher drift values correspond to lower cosine similarity and therefore indicate greater semantic change between stages. Because subsequent analyses report cosine similarity directly, lower similarity values should be interpreted as greater representational drift.

Because each student contributes one essay per stage, embedding comparisons primarily capture assignment-driven variation rather than within-stage stylistic differences. Accordingly, embeddings are used to model trajectories in semantic representation across instructional stages.

This analysis complements the feature-based approach, in which WAT dimensions capture interpretable variation across linguistic functions, while embeddings capture movement through semantic representation space.

3.4 Functional Linguistic Dimension Models

3.4.1 Variance Structure of Functional Writing Dimensions

To characterize the structure of writing development, variance for each dimension was partitioned into student-level, section-level, and residual components using ANOVA-based decomposition ($Y \sim C(student\_id) + C(asu\_class\_id)$).

Results show substantial between-student stability across several dimensions (Table 2). Sophisticated Wording (0.57), Sentence Cohesion (0.54), Word Variety (0.54), and Conversational Writing Style (0.52) exhibit the largest student-level variance components, with Academic Focus (0.47) and Conventional Language (0.41) also showing substantial stability.

In contrast, Development of Ideas (0.24), Word Concreteness (0.32), and Information Density (0.37) show greater residual variance, indicating higher within-student fluctuation across instructional stages. Section-level variance is negligible across all dimensions (≤ 0.01), suggesting minimal structural differences between course sections.

These results indicate that writing development comprises both stable, trait-like dimensions and more plastic, stage-sensitive components, reflecting partially independent subsystems that differ in responsiveness to instructional context.

3.4.2 Developmental Change Across Instructional Stages

Linear mixed-effects models were estimated for each dimension with stage index as a fixed effect and random intercepts and random slopes for instructional stage at the student level, allowing individual variation in both baseline performance and stage-dependent change. Model convergence diagnostics were inspected for all mixed-effects models, and no major convergence issues were observed. For the growth-model analyses, p-values were adjusted using the Benjamini–Hochberg false discovery rate (FDR) procedure ($\alpha = .05$) across linguistic dimensions.

Results reveal systematic but non-monotonic stage-dependent shifts (Figure 1). Conventional Language shows the strongest positive growth ($b = 0.22$, $SE = 0.04$), followed by Academic Focus ($b = 0.11$, $SE = 0.03$) and Information Density ($b = 0.09$, $SE = 0.04$), indicating increasing formalization, informational density, and greater adherence to grammatical and conventional language use across stages.

In contrast, Development of Ideas decreases significantly ($b = -0.15$, $SE = 0.04$), and Word Variety shows a smaller decline ($b = -0.07$, $SE = 0.03$), suggesting a shift from more exploratory writing toward more constrained, concise, and rhetorically focused expression.

Sophisticated Wording, Sentence Cohesion, and Word Concreteness do not exhibit significant linear change, consistent with their higher trait-level stability.

Functional Dimension	b	SE	Adj. p
Development of Ideas	-0.15	0.04	$< .001$
Word Variety	-0.07	0.03	.025
Conversational Writing	-0.01	0.04	.813
Sentence Cohesion	0.03	0.03	.415
Sophisticated Wording	0.04	0.03	.234
Word Concreteness	0.07	0.05	.115
Information Density	0.09	0.04	.010
Academic Focus	0.11	0.03	.001
Conventional Language	0.22	0.04	$< .001$

Table 1: Linear Growth Effects Across Instructional Stages

Dimension	Student	Section	Residual
Soph. Wording	0.57	0.01	0.43
Sent. Cohesion	0.54	0.00	0.46
Word Variety	0.54	0.00	0.46
Conv. Writing	0.52	0.01	0.48
Academic Focus	0.47	0.00	0.53
Conv. Language	0.41	0.01	0.58
Info Density	0.37	0.01	0.62
Word Concr.	0.32	0.01	0.68
Dev. Ideas	0.24	0.01	0.75

Table 2: Variance partitioning of functional writing dimensions.

Overall, these findings suggest stage-dependent reorganization toward academic register, with gains in grammatical control, concision, and focus rather than uniform improvement across all dimensions.

3.4.3 Cross-Dimensional Coupling

To examine dynamic coupling between functional dimensions, lag–difference models were estimated in which stage-to-stage change in a focal dimension (ΔY_t) was regressed on the prior-stage value of another dimension (X_{t-1}). Models were estimated with cluster-robust standard errors at the student level. Based on theoretical expectations regarding register formalization and the relationship between elaboration and academic focus, consistent with prior work contrasting conversational and informational academic discourse (Biber, 2012), three directional pairings were specified for cross-lagged testing; p-values are reported without adjustment given the small number of pre-specified comparisons:

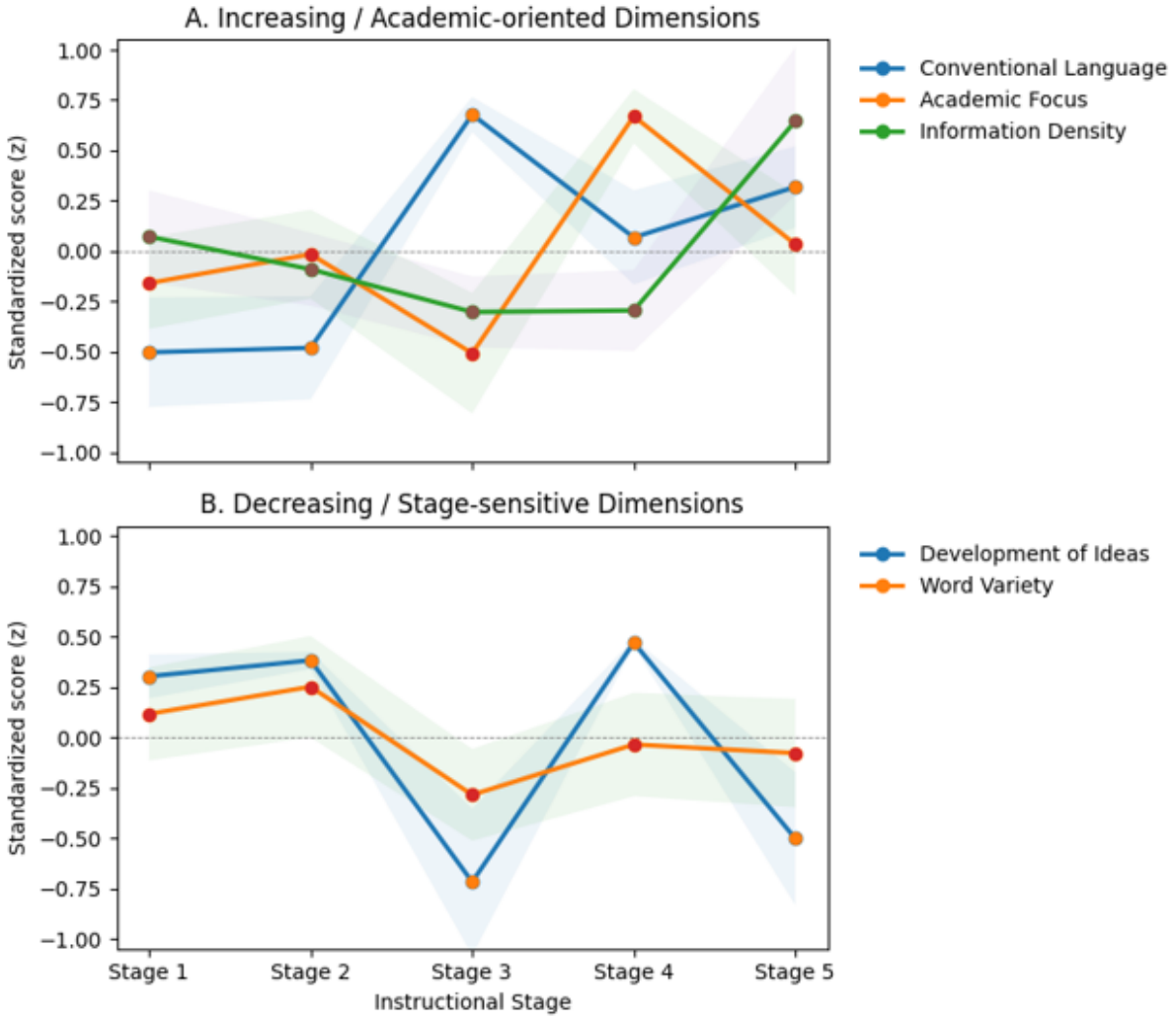


Figure 1: Developmental trajectories of selected WAT dimensions across instructional stages. (A) Increasing dimensions (Conventional Language, Academic Focus, and Information Density) show consistent growth across stages, reflecting movement toward formal and academically oriented writing. (B) Stage-sensitive dimensions (Development of Ideas and Word Variety) exhibit declines and fluctuations, indicating shifts away from exploratory discourse toward more constrained rhetorical structure. Shaded regions indicate variability across students.

$$\Delta Y_t = Y_t - Y_{t-1} = \alpha + \beta X_{t-1} + \varepsilon_t \quad (3)$$

Results indicate significant cross-dimensional interactions across the three pre-specified directional pairings (Table 3). Prior Academic Focus positively predicts change in Conversational Writing ($b = 0.18$, $SE = 0.07$, $p = .012$), while prior Conventional Language predicts increases in Information Density ($b = 0.13$, $SE = 0.06$, $p = .046$), suggesting coordinated and partially reinforcing dynamics across dimensions.

The strongest effect is observed for Development of Ideas, which negatively predicts change in Academic Focus ($b = -0.70$, $SE = 0.11$, $p < .001$), suggesting an inverse relationship between exploratory and more formal writing across stages, although this effect should be interpreted cautiously because the lag–difference specification does not control for prior outcome levels.

Pred.	**Outcome (Δ)**	**b**	**p**
Acad. Focus	Δ Conv. Writing	0.18	.012
Conv. Lang.	Δ Info Density	0.13	.046
Dev. Ideas	Δ Acad. Focus	-0.70	$< .001$

Table 3: Cross-lagged effects on stage-to-stage change.

Overall, these results show that writing development reflects interacting linguistic subsystems under instructional constraints.

Comparison	Cosine Similarity (SD)
Same student essays	0.22 (0.13)
Different student essays	0.23 (0.16)

Table 4: Representational similarity within and between students.

Stage Transition	Cosine Similarity (SD)
Stage 1 → 2	0.15 (0.10)
Stage 2 → 3	0.10 (0.08)
Stage 3 → 4	0.23 (0.10)
Stage 4 → 5	0.28 (0.13)

Table 5: Representational movement between instructional stages.

3.4.4 Embedding-based Results

We first compared semantic similarity between essays written by the same student and those written by different students. As shown in Table 4, similarity was comparable (same-student: $M = 0.22$, $SD = 0.13$; different-student: $M = 0.23$, $SD = 0.16$), with no significant difference ($p = .30$). This indicates that embeddings primarily capture assignment-driven variation. This pattern is consistent with the semantic orientation of SBERT-based representations, which are optimized to encode contextual and topical similarity rather than stable author-specific stylistic identity.

We then examined similarity across consecutive stages for the same student (Table 5). Early transitions show lower similarity (Stage 1→2: $M = 0.15$; Stage 2→3: $M = 0.10$), indicating substantial representational change. Later transitions show higher similarity (Stage 3→4: $M = 0.23$; Stage 4→5: $M = 0.28$), suggesting increased stability in later stages.

Finally, cross-student similarity varies across stages (Table 6), with the highest similarity at Stage 3 ($M = 0.54$) and lower similarity in earlier and later stages (e.g., Stage 2: $M = 0.20$, Stage 5: $M = 0.23$). This pattern indicates that assignment structure shapes the distribution of texts in semantic space.

Taken together, these results show that embeddings capture stage-dependent, non-linear trajectories driven by instructional context, complementing feature-based analyses that reflect stable variation and subsystem dynamics.

Stage	Mean Pairwise Similarity (SD)
Stage 1	0.37 (0.21)
Stage 2	0.20 (0.11)
Stage 3	0.54 (0.12)
Stage 4	0.43 (0.22)
Stage 5	0.23 (0.13)

Table 6: Stage-level cross-student similarity in embedding space.

4 Discussion

4.1 WAT Dimensions and Implications for Computational Language Learning

The variance structure indicates that writing development preserves stable individual differences alongside stage-dependent change. Several dimensions—particularly lexical sophistication, cohesion, and stylistic features such as academic focus and conversational writing style—exhibit strong trait-level stability, suggesting that models of language development should capture persistent individual variation in addition to aggregate trends.

Developmental trajectories are non-uniform. While academic focus, formal language, and information density increase over time, development of ideas and lexical variety decline. This pattern reflects reweighting among partially competing dimensions rather than uniform improvement, highlighting a limitation of computational models that assume monotonic performance scaling.

These patterns emerge within a structured sequence of assignments with distinct rhetorical demands, suggesting that observed trajectories reflect stage-dependent reorganization under instructional and rhetorical constraints rather than isolated developmental progression alone. However, the presence of stable between-student variation and cross-stage dependencies indicates that the results are not solely attributable to assignment effects. Rather than isolating development from instructional context, writing reorganizes across pedagogically sequenced tasks, with different assignments eliciting shifts in linguistic priorities. Although assignments differ in rhetorical demands, they are designed within introductory composition courses to develop transferable knowledge about writing and rhetorical situations, and prior work suggests that writing quality reflects multiple linguistic dimensions that generalize across tasks (Crossley, 2020).

Cross-lagged analyses further show that dimensions interact over time. The negative relation-

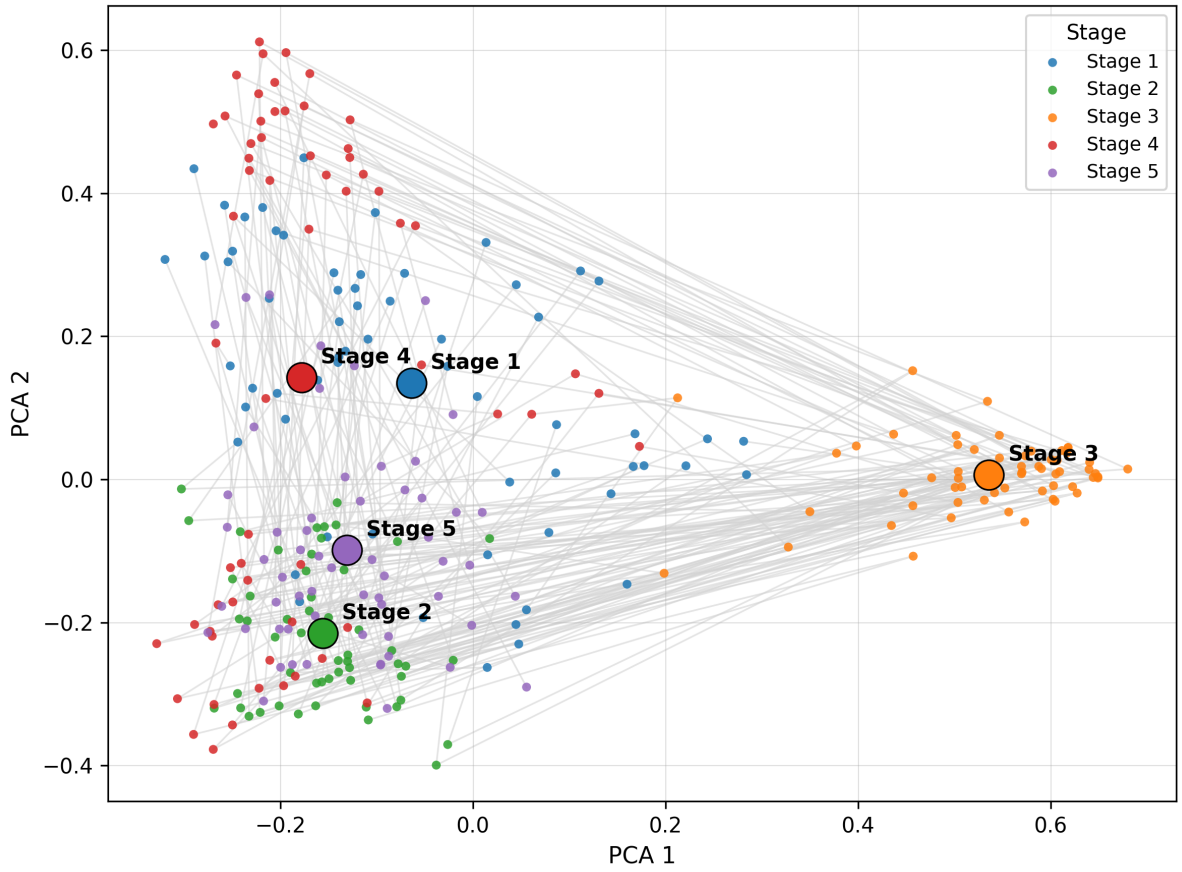


Figure 2: SBERT representation trajectories across instructional stages. Thin gray lines connect successive essays by the same student in PCA-projected embedding space. Colored points represent essays by stage, and larger markers indicate stage centroids. The projection shows substantial overlap and dispersion across stages, with heterogeneous and non-linear trajectories. Certain stages (e.g., Stage 3) occupy relatively distinct regions, consistent with stage-dependent variation observed in similarity analyses.

ship between Development of Ideas and subsequent Academic Focus suggests a possible inverse relationship between exploratory and formal writing, consistent with coordinated adjustments across interacting components rather than independent progression.

Together, these results support a view of writing development as a multidimensional system characterized by stable variation, stage-dependent change, and interactions among competing dimensions (Zhu and Jurgens, 2021). For computational models to approximate human language development, they should account for variability, non-monotonic trajectories, interactions among representational components, and the influence of task structure on observed behavior.

4.2 Embedding-based Analysis

Figure 2 visualizes SBERT-based representation trajectories across instructional stages. The projection shows substantial overlap and dispersion across stages, indicating that semantic representations vary widely across students and assignments. While certain stages, such as Stage 3, occupy relatively distinct regions of embedding space, most stages exhibit considerable overlap, suggesting that assignment-driven semantic structure does not produce clear separation in low-dimensional projections.

Individual trajectories further illustrate that representational change is non-linear and heterogeneous across students, with paths frequently crossing and diverging rather than following a uniform progression. This pattern is consistent with the quantitative results, which show stage-dependent changes in similarity but no strong evidence of stable author-specific clustering.

Embedding-based analyses indicate that semantic representations are primarily shaped by instructional context. The absence of differentia-

tion between same-student and different-student essays suggests that embedding similarity reflects assignment-driven variation rather than stable stylistic differences. At the same time, embeddings capture stage-dependent, non-linear trajectories, with larger representational shifts occurring in earlier stages and greater stability emerging in later stages.

Cross-student similarity patterns further indicate that assignment structure shapes the distribution of texts in representation space, with some stages producing more homogeneous responses than others. Taken together, these findings show that embeddings capture trajectories in semantic space driven by instructional sequencing, complementing feature-based analyses that reflect stable variation and interactions among linguistic dimensions. Writing development can therefore be understood as an interaction between learner-level variation and task-level constraints, with different representational frameworks capturing distinct aspects of this process.

5 Conclusion

This study models writing development as a multidimensional system shaped by stable individual variation and instructional progression across staged assignments. Using interpretable linguistic dimensions, we show that some aspects of writing exhibit strong stability across students, while others change systematically with instructional demands.

Patterns across instructional stages are non-uniform: academic focus, formal language, and information density increase over time, whereas development of ideas and lexical variety decline. Cross-dimensional analyses further reveal interactions among dimensions, including a possible inverse relationship between exploratory and more formal writing.

Embedding-based analyses complement these findings by capturing stage-dependent, non-linear trajectories in semantic representation. Rather than reflecting stable stylistic similarity, embeddings primarily encode assignment-driven variation, with larger shifts in early stages and greater stability later.

Together, these results characterize writing development as coordinated change across interacting dimensions under instructional constraints.

Limitations

A key consideration is that instructional stages correspond to different assignment types with distinct rhetorical demands that are intentionally sequenced within the curriculum. As a result, observed changes reflect both developmental processes and structured task progression. This design provides a controlled instructional framework for examining how writing reorganizes across pedagogically defined stages, but it does not fully isolate development from task-specific effects. Future work could extend this approach by examining development within repeated or more tightly controlled task settings.

Because lag–difference models did not include lagged outcome controls, cross-dimensional effects should be interpreted as suggestive dependencies rather than strong causal or mechanistic relationships.

Additionally, the dataset is drawn from a single institutional context and course, with 62 students contributing repeated observations across instructional stages. Although the longitudinal design provides multiple observations per student, the modest sample size may limit statistical power and the stability of more complex model estimates. Findings may therefore not generalize to other instructional settings, disciplines, or curricula.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305N210041 and R305T240035 to Arizona State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Douglas Biber. 2012. Register as a predictor of linguistic variation. *Corpus linguistics and linguistic theory*, 8(1):9–37.

Scott A Crossley. 2020. Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3):415–443.

Scott A Crossley and Minkyung Kim. 2022. Linguistic features of writing quality and development: A longitudinal approach. *The Journal of Writing Analytics*, 6(1):59–93.

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.

James Fiacco, Shiyan Jiang, David Adamson, and Carolyn Rosé. 2022. Toward automatic discourse parsing of student writing motivated by neural interpretation. In *Proceedings of the 17th workshop on innovative use of NLP for building educational applications (BEA 2022)*, pages 204–215.

Larissa Goulart and Margaret Wood. 2021. Methodological synthesis of research using multi-dimensional analysis. *Journal of Research Design and Statistics in Linguistics and Communication Science*, 6(2):107–137.

Danielle S McNamara, Tracy Arner, Elizabeth Reilley, Paul Alvarado, Chani Clark, Thomas Fikes, Annie Hale, and Betheny Weigele. 2022. The asu learning at scale (asu l@ s) digital learning network platform. *Grantee Submission*.

Danielle S. McNamara, Andrew Potter, Zeinab Serhan, Manmeet Singh, Nishad Patne, Tracy Arner, Rod D. Roscoe, Laura K. Allen, and Scott A. Crossley. 2026. The writing analytics tool: A learning engineering approach to designing ai-supported writing instruction. In *Proceedings of the Learning Engineering Research Network Convening (LERN 2026): From Insights to Implementation, Learning Engineering in Action*. LERN Convention Proceedings.

Andrew Potter, Zeinab Serhan, Nishad Patne, Püren Öncel, Ishrat Ahmed, Tracy Arner, Rezwana Islam, Rod Roscoe, Laura Allen, Scott Crossley, and Danielle McNamara. 2026. Human-ai collaboration for qualitative analysis in participatory design: Refining the writing analytics tool. *Journal of Educational Data Mining*, 18(1):113–155.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268.

Jian Zhu and David Jurgens. 2021. Idiosyncratic but not arbitrary: Learning idiolects in online registers reveals distinctive yet consistent individual styles. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 279–297.

L1 Influence in L2 Language Models: A Human-centric Approach

Laura Barbenel[1,2*] **Lily Goulder**[1,2†] **Aoife O'Driscoll**[1,2†] **Suchir Salhan**[1,2]
Catherine Arnett[3] **Andrew Caines**[1,2] **Paula Buttery**[1,2]
[1] ALTA Institute [2] Department of Computer Science & Technology, University of Cambridge
[3] EleutherAI
Correspondence: lgb35@cam.ac.uk

Abstract

Language learners typically exhibit first language (L1) influence in their written second language (L2) production. We investigate whether similar patterns emerge in L2 language models (L2LMs), which are typically assessed on task-based benchmarks rather than on language use. We evaluate the use of Native Language Identification (NLI) as a method for detecting whether L2LMs exhibit human-like L1 influence. Using existing learner corpora and our novel L2 English dataset, we identify the conditions that yield the highest NLI accuracy, and show that text length but not proficiency affects performance. We then apply NLI to L2LM-generated text under various instruction-tuning and prompting conditions. We find that instruction tuning on human learner essays yields high NLI accuracy (~90%) and is necessary for detectable L1 influence. Whilst NLI accuracy is similar for L2LM and human essays, human evaluation shows that LM-generated L1 influence remains distinguishable from human writing.

1 Introduction

Multilingual language models (LMs) are trained with shared parameters across languages, facilitating crosslingual generalisation (Conneau et al., 2020). Yet bilingual and multilingual models often underperform compared to monolingual systems (Zhou and Matusevych, 2025; Xu et al., 2025; Salhan et al., 2026). However, these findings are based on evaluations that define linguistic competence primarily in terms of benchmark accuracy, such as controlled grammaticality judgement tasks (e.g., BLiMP; Warstadt et al., 2020), which quantify task performance but do not capture an LM's linguistic *behaviour* (Hagendorff et al., 2023), or whether its crosslingual generalisations resemble patterns observed in human second language acquisition (SLA). Consequently, it remains unclear whether bilingual LMs exhibit human-like behaviour in second language (L2) production.

In human SLA, a learner's first language (L1) systematically influences written L2 production (Swan and Smith, 2001; Gonzalez-Torres and Mayo, 2025; Shatz, 2017), with errors emerging when L1 and L2 patterns diverge (Selinker, 1969). We investigate whether second language LMs (L2LMs)—models designed to simulate SLA (Aoyama and Schneider, 2024)—can generate human-like L2 text. Human-likeness is defined as the presence of L1-influenced patterns observed in human L2 writing. Our study focuses on seven L1s (Spanish, French, German, Polish, Turkish, Arabic, Mandarin Chinese) with English as the L2.

We evaluate L2LMs using Native Language Identification (NLI), a method for detecting a writer's L1 from L2 text (Tetreault et al., 2013); this application allows us to assess whether L2LM outputs exhibit human-like L1 influence. To our knowledge, this is the first time that NLI has been applied to LM-generated text. First, we identify the conditions under which NLI is most accurate on human-written text. Based on our findings, we use the best-performing NLI method to evaluate L2LM-generated text. We address two research questions:

RQ1: What are the optimal conditions for NLI detection of L1 influence?

We compare the NLI capabilities of GPT-4 and GPT-5 under two prompting conditions (explanation vs. no explanation). We also examine text-level factors such as essay length and learner proficiency. To disentangle these typically confounded effects, we create a length-controlled L2 learner dataset spanning three proficiency levels (beginner, intermediate, advanced) and seven L1 backgrounds, with matched essay prompts. This enables the first

[*] Corresponding author
[†] Equal contribution

Proceedings of the 1st Workshop on Computational Developmental Linguistics (CDL), pages 92–116

systematic investigation of how text length and proficiency independently affect NLI performance. Our results indicate that GPT-5 and prompting for explanations provide no significant advantage over GPT-4 without explanation. Crucially, we find that NLI performance is strongly affected by text length, but not learner proficiency, validating it as a robust detection tool across proficiency levels.

RQ2: Can L2LMs simulate human-like L2 writing?

We investigate factors that shape L1 influence in L2LM outputs, including instruction-tuning, prompt style, and the quantity of L2 pretraining data. We compare base L2LM essays with those generated under three instruction-tuning conditions: English-only, English and L1, and L1-specific L2 English, a dataset created for this study. We find that LM-generated essays are accurately classified by NLI only when instruction tuned on L1-specific L2 English. In this instruction-tuning condition, neither L2 pretraining data nor prompt style manipulations affected NLI accuracy.

To address human-likeness, we compare LM-generated essays with human-written L2 essays matched for L1, prompt, and text length. Whilst NLI accuracy is equally high on LM and human essays, analyses of named entities show that they affect NLI-detectable L1 influence in LM-generated essays but not in human-written essays. Finally, human evaluation reveals that, despite high NLI accuracy, L1 influence in L2LM-generated text is distinguishable from human writing, indicating that L2LMs do not simulate human-like written L2 production.

Contributions Our study contributes: (1) an identification of conditions that optimise NLI performance, including effects of text length and proficiency; (2) a novel human L2 corpus of prompt-matched, length-controlled essays from seven L1s and three proficiency levels; (3) a novel L1-specific L2 human learner instruction-tuning dataset, and demonstration of its necessity for detectable L1 influence in L2LM outputs; (4) a comparison of LM-generated and human-written L2 texts, marking the first application of NLI to LM-generated text; (5) human evaluation showing that high NLI accuracy does not guarantee human-like L2 production.

Models, tokenizers, and datasets are available on Hugging Face.[1] Pretraining, instruction tuning, NLI, and prompting code is available on GitHub.[2]

2 Related Work

2.1 L2LMs as human proxies

There are several clear limitations to using LMs as proxies for human learning (see Cuskley et al., 2024; cf. Salhan et al., 2025). Firstly, LMs are trained on orders of magnitude more data than humans. Although smaller human-scale models may be more appropriate for cognitive comparison (Wilcox et al., 2025), they diverge from human learners in fundamental ways. Whilst efforts such as the BabyLM challenge aim to reduce this gap (Warstadt et al., 2023), the nature of the input remains fundamentally different: LMs are trained on text, whereas human learners receive multimodal, interactive input in an *embodied* manner (Cuskley et al., 2024). Additionally, learning in LMs relies on backpropagation, a learning mechanism that is biologically implausible (Lillicrap et al., 2020). Moreover, LMs may succeed on some tasks but fail to replicate human-like processing (see Cai et al., 2024; Binz and Schulz, 2023). High performance may reflect architectural flexibility rather than human-like representations, as models perform equally well when trained on unnatural inputs such as reversed text (Luo et al., 2024).

Despite differences between humans and LMs, prior work suggests that L2LMs can simulate aspects of human L2 behaviour. For example, structural priming effects in L2LMs are weaker for typologically distant L1-L2 pairs (Arnett et al., 2025), mirroring human cross-linguistic priming (Hartsuiker et al., 2004; Bernolet et al., 2007). For L2 competence, Yadavalli et al. (2023) provide indirect evidence of L1 influence: poor performance on specific L2 constructions in BLiMP is attributed to L1 transfer. For L2 production, LLMs can generate L1-influenced dialogue when prompted with human L2 data and descriptions of L1-specific grammatical traits (Gao et al., 2025). Since these L1 effects are induced and the models are not L2LMs, it remains unclear whether human-like L1 influence in L2 production can emerge naturally. Building on this, we evaluate whether L2LMs can *simulate* human-like L2 text.

[1] https://huggingface.co/ALTACambridge

[2] https://github.com/suchirsalhan/l1-influence-in-l2lms

2.2 L1 influence in L2 production

Human SLA exhibits characteristic patterns of L1 influence. Language transfer—the effect of the L1 on the L2—can facilitate learning via typological similarities or inhibit it due to differences, causing systematic error patterns (Swan and Smith, 2001; Gonzalez-Torres and Mayo, 2025). L1 influence can also appear subtly, such as in the overuse of L1–L2 congruent structures or avoidance of L1-absent structures (Berti et al., 2023; Zomer and Frankenberg-Garcia, 2021). Since different L1s produce unique influence patterns (inter-group heterogeneity; Jarvis, 2000), a learner's L1 can be inferred through detection-based approaches (Jarvis and Crossley, 2012). This motivates the machine learning task of NLI, which can detect "subtle and unpredicted" L1 transfer (Yang et al., 2025).

2.3 Native Language Identification

NLI involves determining an author's L1 from text written in an L2. Traditional approaches employ feature-based methods (e.g., Tetreault et al., 2013), which can reach ~80% accuracy but require extensive feature engineering. In contrast, LLMs have established a new state of the art, exceeding 90% accuracy (e.g., Zhang and Salle, 2023; Ng and Markov, 2025).

In traditional NLI methods, classification is typically easier for low-proficiency essays (Kyle et al., 2015). The effect of learner proficiency on LLM-based methods has not yet been tested. However, it has been shown that NLI accuracy improves with increased text length (Zhang and Salle, 2023; Nicholls and Alperin, 2025). Crucially, in L2 writing, text length and proficiency are strongly correlated (Martínez, 2018), and the individual effect of each factor, whilst controlling for the other, has not yet been systematically investigated.

3 Methodology

To investigate whether L2LMs exhibit human-like L1 influence (RQ2), we evaluated their outputs using NLI under the optimal conditions identified in RQ1.

3.1 RQ1

To identify optimal NLI conditions, we conduct two experiments. Experiment 1 compares model (GPT-4 vs. GPT-5) and prompting (explanation vs. no explanation) conditions to identify the best-performing configuration for L1 detection. Experiment 2 examines text-level factors, testing the effect of learner proficiency whilst controlling for text length to isolate its impact on NLI performance.

Experiment 1 Following Zhang and Salle (2023), the model performs NLI as a closed-set, zero-shot task (see Figure 8, Appendix B). Four model conditions were compared: two models (GPT-4 and GPT-5[3]) under two prompting conditions – requesting explanation vs. no explanation. Whilst GPT-4 currently achieves state-of-the-art performance on NLI benchmarks (Zhang and Salle, 2023; Ng and Markov, 2025), GPT-5 has not yet, to the best of our knowledge, been evaluated for this task. Explanation prompting is motivated by evidence that generating explanations can improve LLM performance (Dhaini et al., 2025). NLI was run five times for each model condition to ensure reliability.

We used a subset of the Write & Improve Corpus 2024 (W&I; Nicholls et al., 2024), which includes L1 and CEFR[4] proficiency metadata. The subset comprised 147 essays (21 per L1; seven per CEFR level A, B, and C, corresponding to beginner, intermediate, and advanced, respectively) with a mean length of 884 (SD=505) characters.

Experiment 2 We examined the effect of proficiency on NLI accuracy whilst controlling for text length. As prior work suggests that text length affects accuracy (see Section 2.3), it was crucial to isolate the effects of each factor and ensure that NLI performance is comparable across beginner, intermediate, and advanced proficiency levels. We included two text length controls: (1) W&I essays were artificially truncated to the median length of A-level essays (291 characters), with shorter essays (n = 24) included without modification; (2) 466 longer essays (800-1,200 characters) were from our own L2 dataset (see Section 3.3). For comparison, the effect of proficiency without controlling for text length was also evaluated using the original W&I subset. The NLI procedure followed the design of Experiment 1.

3.2 RQ2

We investigated whether L2LMs can simulate human-like L2 writing by examining the effects of four independent variables: instruction-tuning condition, amount of L2 pretraining data received

[3] Accessed Jan-Feb 2026.

[4] Common European Framework of Reference for Languages (CEFR) rating.

(25%, 50%, or 75% of total L2 tokens), prompt style, and prompt topic.

3.2.1 L2LM Training

Pretraining Following Arnett et al. (2025), we pretrained seven 250M-parameter GPT-2 L2LMs on 5B tokens, one per L1, with English as the L2. Each model has 24 layers, 14 attention heads, a hidden size of 896, a 512-token context window, and a 50k-token vocabulary (see Appendix A.1).

The pretraining dataset consisted of L1 and L2 data in a 2:1 ratio. L1 data (3.33B tokens) were sourced from CulturaX (Nguyen et al., 2024), and L2 data (1.67B tokens) from the English subset of FineWeb-Edu (Lozhkov et al., 2024). We trained 50k-vocab bilingual (L1-English) byte-level BPE tokenizers on 2M sentences with a 1:1 L1:L2 mix, and pretokenized all training data into 512-token sequences.

Pretraining followed a two-phase schedule. In Phase 1 (0-50%), the model was exposed exclusively to L1 (2.5B tokens). Phase 2 (51-100%) interleaved the L2 data (1.67B) with the remaining L1 data (0.83B) at a 2:1 ratio.

Instruction tuning We investigated the effect of instruction tuning on L1 influence, with the base (non-instruction-tuned) model as a baseline. Full-weight instruction tuning was applied under three dataset conditions: (1) English-only, (2) English and L1 (1:1), and (3) L1-influenced L2 English. The English-only dataset comprised 30,000 generic QA pairs from Alpaca English (Taori et al., 2023). The English + L1 dataset also contained 30,000 QA pairs, split equally between Alpaca English and its translation for each L1 (see Appendix A.2 for dataset references). We created seven L1-specific L2 English datasets from the EFCAMDAT (Geertzen et al., 2013) and W&I (Nicholls et al., 2024) learner corpora, each containing 10,000 prompt-essay pairs, except Polish (n=370) and Turkish (n=4,847) due to limited corpus data (see Appendix A.2 for detailed dataset composition).

To examine the effect of L2 data exposure, instruction tuning was conducted at one of three pretraining checkpoints: after 25%, 50%, and 75% of total L2 tokens. For ease of comparison with human responses, we refer to these as beginner, intermediate, and advanced proficiency levels, but we do not imply direct equivalence.

In total, 84 model versions were created, covering all combinations of seven L1s, four model variants (one base and three instruction-tuned), and three pretraining checkpoints.

3.2.2 Prompt engineering

As monolingual and bilingual prompting have been shown to affect LLM output (Yuan et al., 2025), we compare three prompting styles: basic, mixed-language, and 'language-to-thought' (L2T). Basic prompts align with human L2 writing assessment tasks. Mixed-language prompts combine L1 and L2 by translating sections of the basic prompts (Srivastava and Singh, 2021; Singh et al., 2024; Lin et al., 2022). L2T prompts instruct the model to 'think' in the L1 before responding in English (see Kang and Kim, 2025), and are otherwise identical to the basic prompts.

```
Prompt topic: Holidays
Prompt style: Mixed language
Prompt:
Tu amiga, Sophia, has emailed you about her recent
holiday. Write an email to her, describiéndole tus
vacaciones favoritas.
Write around 150 words.
```

Figure 1: Example of a mixed-language prompt (Spanish L1).

We created nine distinct prompts by combining three styles (basic, mixed-language, and L2T) with three topics (hobbies, holiday, family). An example is provided in Figure 1 (see Appendix A.3 for all prompts).

Each model version was prompted twice on each of the nine prompts, resulting in a total of 1,512 LM-generated essays. Outputs were constrained to 250-300 tokens (~800-1,200 characters).

3.3 Novel L2 learner dataset

We created a novel dataset of learner English essays by collecting responses to nine prompts from seven L1 backgrounds across three proficiency levels, controlling for text length (RQ1). This design enabled direct comparison with model-generated text by matching prompts between human participants and LMs (RQ2). For each proficiency level and L1, at least six participants were recruited. In total, the dataset comprised 466 essays. Full ethical approval was granted by the Department of Computer Science and Technology at the University of Cambridge.

Participants 174 participants (M age=30; 42.5% female) were recruited via Prolific. The task took

approximately 30 minutes, and participants were compensated at the local living wage. Informed consent was obtained in the participant's native language.

Design and procedure Each participant responded to three prompts (each topic and style were presented exactly once; see Appendix B.3). Prompt sequences were pseudo-randomised to counterbalance the nine prompts across L1-proficiency subgroups, minimising dependencies between prompts, and ensuring that each prompt was answered at least twice per subgroup. Essay length was constrained to 800-1,200 characters. Participants were instructed not to use external resources and were informed that the task was not a test. Essays were manually screened for signs of generative AI usage.

3.4 Evaluation

Model-generated and human-written texts were evaluated with NLI using the GPT-4 model (no explanation) (see Section 4.1). In line with Zhang and Salle (2023), if the model predicted an L1 outside the seven L1 classes, the model was re-prompted until one of the known L1s was predicted.

LLMs may rely on named entities as superficial cues to the writer's L1 (Uluslu et al., 2025). Using Named Entity Recognition (NER), we replaced any entities present in human-written and LM-generated essays with neutral English ones (see Appendix A.4). NLI was performed on both the original and NER-normalised texts.

To evaluate the use of NLI as a metric of human-likeness, three computational linguists judged whether texts were human-written or LM-generated. Each evaluator reviewed 42 texts (21 human, 21 LM), balanced across L1, prompt style, and instruction-tuning condition (126 essays in total). We also assessed the potential effect of LM generation artefacts on judgements by manually editing the essays (e.g., removing unusual characters) provided to one of the evaluators (see Appendix A.4).

4 Results: RQ1

4.1 Experiment 1

Experiment 1 tested NLI performance across four model conditions. Table 1 summarises accuracy by model and prompting condition. No significant differences were found across the five NLI runs for any condition (Table 9, Appendix B).

	explanation	no explanation
GPT-4	83.3	81
GPT-5	79.7	80.4

Table 1: NLI accuracy (%) per model condition (averaged across five iterations).

All four model conditions performed significantly above chance (1/7; Wilcoxon signed rank test, Holm-adjusted $p < .001$). Accuracy did not differ between model conditions (Friedman test, $\chi^2(3) = 3.82, p = .28$). By L1, GPT-5 showed higher accuracy for Arabic ($\chi^2(3) = 11.54, p < .01$), but post-hoc pairwise comparisons were not significant after Holm correction ($p \geq .12$). No other L1 differences were observed. Confusion matrices are shown in Figure 9 (Appendix B). Kruskal-Wallis tests confirmed no significant L1 differences within any model (Table 9, Appendix B).

Based on these results, the GPT-4 no-explanation condition was selected for NLI, as GPT-5 showed no advantage, and explanation prompts provided no benefit, with occasional hallucinations (see Figure 10, Appendix B).

4.2 Experiment 2

We examined the effect of proficiency on NLI accuracy whilst controlling for text length. We report only the results of the GPT-4 no-explanation condition (see Appendix B for other conditions).

Original W&I subset NLI accuracy was 81%, noticeably lower than the state-of-the-art performance (91.7%) reported by Zhang and Salle (2023) on TOEFL11. This is partly attributable to the shorter essays in our dataset (mean 884 vs. 1,785 characters in Zhang and Salle, 2023). Correctly classified W&I essays were significantly longer than misclassified ones (Figure 2; see also Table 10 and Figure 11, Appendix B).

Proficiency and text length are highly correlated in this dataset (Spearman's $\rho(145) = .87, p < .001$). Proficiency significantly affected NLI accuracy (Kruskal-Wallis: $p < .01$), with A-level essays classified significantly less accurately than C-level essays (post-hoc Wilcoxon, $p < .01$; Table 11, Appendix B).

Truncated W&I subset Truncating essays to ≤291 characters reduced overall NLI accuracy to 67%, which remained significantly above chance ($p < .001$). A Kruskal-Wallis test revealed no sig-

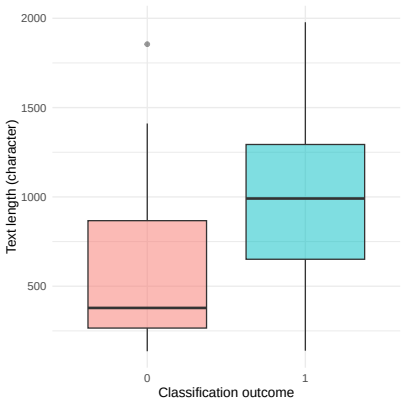
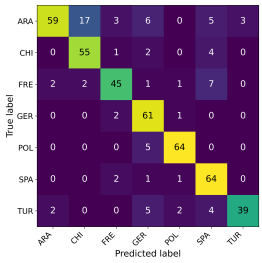


Figure 2: Boxplot of text length by NLI classification accuracy (averaged over five runs). Texts with mean accuracy <1 are classified as 0 (incorrect); mean accuracy = 1 are classified as 1 (correct).

nificant differences in accuracy between A-, B- and C-level essays ($\chi^2(2) = 0.45$, $p = .80$).

Length-controlled L2 dataset NLI accuracy was 83% (95%CI[79.33, 86.34], $p < .001$) on our novel dataset (M = 906 ± 109 characters). Per-L1 accuracy (Figure 3a) was significantly above chance (see Table 3b). A generalised linear mixed model (GLMM) was fitted with L1, proficiency, prompt style, prompt topic, and NER condition as fixed effects, and user ID as a random intercept (see Table 14, Appendix B.3 for model summary). Crucially, proficiency had no significant effect on NLI accuracy ($p \geq .09$).



(a) Confusion matrix

L1	F1	P	R	*p*
Spanish	0.84	0.76	0.94	<.001
French	0.81	0.85	0.78	<.001
German	0.84	0.75	0.95	<.001
Polish	0.93	0.93	0.93	<.001
Turkish	0.83	0.93	0.75	<.001
Arabic	0.76	0.94	0.63	<.001
Chinese	0.81	0.74	0.89	<.001

(b) Precision, recall, and F1

Figure 3: NLI performance on human-written essays: (a) confusion matrix; (b) per-L1 precision (P), recall (R), and F1 scores. p-values are from one-tailed binomial tests.

Overall, Experiment 2 showed that, when text length was controlled, proficiency had no significant effect on NLI accuracy, both for short (291 characters) and longer (800-1,200 characters) essays, validating GPT-4 as a detection tool across proficiency levels.

5 Results: RQ2

5.1 Effects of instruction tuning

We evaluated the L2LM generations under three instruction-tuning conditions, as well as from the base models as a baseline.

First, the base models were tested on BLiMP (Warstadt et al., 2020) and MultiBLiMP (Jumelet et al., 2026) (see Figure 13, Appendix C.1). MultiBLiMP results suggest that our training regime did not lead to catastrophic forgetting. BLiMP scores increased as L2 data quantity increased. The seven models (one per L1) did not differ significantly in BLiMP and MultiBLiMP scores (see Table 15, Appendix C.1).

5.1.1 Base model

Mean NLI accuracy was 23.5% (95%CI[19.36, 28.15]), remaining significantly above chance ($p < .01$). However, this effect was driven by Chinese, where high recall (0.91) but low precision (0.16) resulted in many false positives (see Table 2; Figure 4a). NLI accuracy was not above chance for any other L1.

L1	F1	P	R	*p*
Spanish	0.22	0.38	0.15	.51
French	0.34	0.75	0.22	.08
German	0.26	0.40	0.19	.24
Polish	0.08	1.00	0.04	1.0
Turkish	0.04	1.00	0.02	1.0
Arabic	0.23	1.00	0.13	.67
Chinese	0.27	0.16	0.91	<.001

Table 2: NLI precision (P), Recall (R), and F1 scores per L1 background for the base model. p-values are from one-tailed binomial tests.

5.1.2 Comparing instruction-tuning conditions

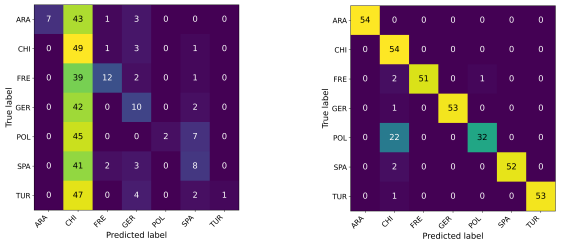
A GLMM was fitted with the fixed effects of L1, instruction-tuning condition, proficiency, prompt style, prompt topic, and NER condition, including the two-way interactions between instruction tuning and NER, prompt style and prompt topic, and instruction tuning and prompt style. Model version was included as a random intercept. Model summary is shown in Table 3 (full results in Table 16, Appendix C).

Baseline accuracy was below chance ($p < .001$). Relative to this baseline, L1 significantly affected performance: Polish, Turkish, and Arabic decreased accuracy, whilst the significant advan-

Term	Estimate	Std. Error	z value	p
(Intercept)	-2.92	0.48	-6.07	<.001
IT: English-only	-0.16	0.52	-0.31	.76
IT: English + L1	0.52	0.48	1.08	.28
IT: L2 English	6.75	0.51	13.33	<.001
NER	0.03	0.26	0.13	.90
Style: L2T	0.10	0.48	0.22	.83
Style: Mixed	2.75	0.43	6.43	<.001
Topic: Hobbies	0.64	0.31	2.04	<.05
Topic: Holiday	0.69	0.31	2.19	<.05
IT: English-only * NER	0.04	0.38	0.12	.91
IT: English + L1 * NER	-0.12	0.35	-0.34	.73
IT: L2 English * NER	-3.37	0.40	-8.35	<.001
Style: Mixed * Topic: Hobbies	-1.03	0.39	-2.62	<.01
IT: English-only * Style: L2T	1.34	0.56	2.41	<.05
IT: English-only * Style: Mixed	-1.22	0.53	-2.33	<.05
IT: L2 English * Style: Mixed	-2.22	0.44	-5.00	<.001

Table 3: Model summary. The intercept is set to base model, Spanish L1, basic prompt style, family topic, and original essays (no NER).

tage of Chinese L1 is attributable to false positives (see Table 2). Proficiency had no significant effect. Neither English-only nor English + L1 instruction-tuning conditions significantly affected accuracy ($p > .05$). In contrast, L2 English instruction tuning significantly improved NLI accuracy ($p < .001$). Whilst NER had no main effect ($p = .90$), it significantly dampened the improvement from L2 English instruction tuning ($p < .001$). Prompt topic influenced accuracy: essays on hobbies and holiday topics were classified more accurately than the baseline ($p < .05$). Prompt style effects depended on both topic and instruction-tuning condition. Mixed-language prompting improved accuracy in the base model (see Appendix C.3). However, this benefit was attenuated for the hobbies topic ($p < .01$), as well as under English-only ($p < .05$) and L2 English ($p < .001$) instruction-tuning conditions. L2T prompts had no significant main effect but significantly increased accuracy under English-only instruction tuning ($p < .05$).



(a) Base model (b) L2 English IT

Figure 4: Confusion matrices for NLI on essays generated by the base model and L2 English instruction-tuning model conditions.

Overall, the L2 English instruction-tuning condition resulted in the highest NLI accuracy. Neither the English-only nor English and L1 instruction-tuning conditions improved NLI performance relative to the base model (see Appendix C for detailed results).

5.1.3 L2 English instruction tuning

Table 4 reports NLI precision and recall for original and NER-normalised essays. On the original essays, overall NLI accuracy was 92.33% (95% CI [89.2%, 94.80%]). Classification accuracy was significantly above chance (1/7) for all L1s ($p < .001$). Precision was near ceiling for all L1s except Chinese (0.65), reflecting false positive predictions (see Figure 4). Recall was near ceiling (0.94-1.00) for all L1s except Polish (0.59), likely due to its smaller intruction-tuning dataset (n = 370 examples; Section 3.2.1). Accuracy did not differ significantly by prompt style ($\chi^2(2) = 5.45, p = .07$) or by prompt topic once accounting for interactions between prompt topic and style (see Table 20, Appendix C.6).

NER normalisation reduced accuracy to 55.82% (95% CI [50.65%, 60.90%]) but remained significantly above chance overall ($p < .001$) and for all L1s except Polish ($p = 1.0$), likely due to the small size of the instruction-tuning dataset (Section 3.2.1). Per-L1 comparisons indicated that NER normalisation significantly reduced NLI accuracy for all L1s ($p < .001$) except Spanish ($p = .13$) (Table 21, Appendix C.6).

L1	**Learner IT (original)**				**Learner IT (NER)**			
	F1	**P**	**R**	***p***	**F1**	**P**	**R**	***p***
Spanish	0.98	1.00	0.96	<.001	0.77	0.68	0.89	<.001
French	0.97	1.00	0.94	<.001	0.72	0.97	0.57	<.001
German	0.99	1.00	0.98	<.001	0.77	1.00	0.63	<.001
Polish	0.73	0.97	0.59	<.001	0.04	0.25	0.02	1.0
Turkish	0.99	1.00	0.98	<.001	0.67	0.83	0.56	<.001
Arabic	1.00	1.00	1.00	<.001	0.41	1.00	0.26	<.05
Chinese	0.79	0.65	1.00	<.001	0.44	0.28	0.98	<.001

Table 4: Precision (P), Recall (R), and F1 scores per L1 background for learner essay instruction-tuning condition: original and NER-normalised essays.

5.2 Comparing human-written and LM-generated texts

To compare NLI accuracy on human-written and LM-generated essays (from the L2 English instruction-tuning condition), a GLMM was fitted with the fixed effects of animacy (human vs.

LM), L1, NER, prompt style, prompt topic, and proficiency. We included interactions between animacy and L1, animacy and NER, and animacy and prompt topic. User ID (model version/participant) was included as a random intercept. Model summary is in Table 5 (full results in Table 22, Appendix C.7).

Term	Estimate	Std. Error	z value	p
(Intercept)	4.13	0.76	5.45	<.001
L1: Turkish	-2.48	0.79	-3.12	<.01
L1: Arabic	-3.25	0.72	-4.54	<.001
Animacy: LM	1.46	1.34	1.09	.28
NER	0.05	0.22	0.23	.82
Animacy: LM * L1: Polish	-5.33	1.74	-3.07	<.01
Animacy: LM * NER	-3.94	0.46	-8.66	<.001
Animacy: LM * Topic: Hobbies	0.89	0.40	2.24	<.05
Animacy: LM * Topic: Holiday	1.00	0.41	2.46	<.05

Table 5: Model summary. The intercept is set to human-written, Spanish L1, basic prompt style, family topic, and original essays (no NER).

Baseline accuracy was significantly above chance ($p < .001$). Turkish and Arabic L1 significantly decreased NLI accuracy ($p < .01$), although performance remained significantly above chance (Table 3b). There were no significant main effects of animacy, NER, prompt style, prompt topic, or proficiency ($p < .05$). However, a strong negative interaction between animacy and NER was found ($p < .001$), indicating that NER significantly reduced NLI accuracy for LM-generated essays but not for human-written essays. The interaction between animacy and L1 was significant for Polish ($p < .01$), with LM-generated essays being classified less accurately than human essays. Additionally, LM-generated essays were classified more accurately than human essays for hobbies and holiday topics ($p < .05$).

5.2.1 Human evaluation

Mean evaluator accuracy was 87.3%. Classification accuracy did not significantly differ between evaluators (McNemar's test, $p = .61$), and removing unusual characters from LM outputs had no effect ($p = 1.0$). Although evaluators misclassified human-written texts as LM-generated three times more often than the reverse, this difference was not statistically significant ($p = .08$). No other conditions significantly affected evaluation outcomes (see Appendix C.8).

6 Discussion

LLM-based NLI methods reliably detect L1 influence in L2LM outputs; equally, human evaluation consistently distinguished LM-generated essays from human L2 production, indicating that NLI detectability should not be interpreted as a measure of human-likeness.

GPT-4 (without explanation prompting) was selected as our NLI tool, as GPT-5 and/or prompting for explanation offered no advantage. NLI successfully detected L1 influence across all proficiency levels once text length was held constant.

In L2LM-generated text, L1 influence was absent in base model outputs and in generic (English-only or English and L1) instruction-tuning conditions. As we did not find evidence of catastrophic forgetting, these results indicate that L1 exposure alone is insufficient to produce detectable L1 influence in L2LM output. L1 influence emerged only in essays generated by LMs that were instruction-tuned on L1-specific L2 English datasets. Notably, these instruction-tuning datasets were only a third of the size of the other instruction-tuning datasets, suggesting that instruction-tuning dataset specificity, rather than quantity, drove L1 influence.

Within L2 English IT, L1 influence was consistently detected across model checkpoints, showing that the quantity of L2 pretraining data had no effect on L1 influence. Additionally, prompting conditions did not affect NLI accuracy: L1 influence was present both with and without explicit L1 elicitation.

NLI accuracy was comparable for human-written and LM-generated essays from the L2 English instruction-tuning condition. For these L2LMs, NER normalisation reduced NLI accuracy, yet performance remained above chance, highlighting that L1 influence went beyond superficial cues. Notably, the adverse effect of NER normalisation on NLI performance was not observed for human-written essays. Moreover, human evaluation clearly distinguished LM-generated from human-written essays.

Additionally, Chinese was overpredicted as the L1 across all LM outputs, but this bias was absent from human essays, suggesting that it stems from pretraining data artefacts rather than limitations of the NLI method. Overall, whilst using NLI as a metric suggests that L1 influence in human and LM essays may be comparable, further analyses reveal that L2LM production remains distinctly non-human-like.

7 Conclusion

Our findings show that L2LMs produce L1 influence only when instruction-tuned on L1-specific L2 English. However, as this L1 influence did not resemble human L2 production, NLI detectability ultimately should not be used as a measure of human-likeness. Overall, L2LM outputs remain distinctly non-human-like, limiting their use as proxies for human L2 learners.

Limitations and Future Directions

NLI accurately detected L1 influence in L2LM-generated essays; however, its precise nature remains unclear. Future work should aim to elucidate the properties of these L1 effects, perhaps by incorporating linguistically motivated fine-grained analyses. Additionally, Chinese false positives in NLI were prevalent across all L2LM essays regardless of instruction-tuning condition. We hypothesise that more rigorous pretraining data cleaning may mitigate this overprediction bias. Furthermore, although the combination of L1-specific pretraining and instruction tuning produced detectable L1 influence, pretraining alone was insufficient. It remains unclear whether instruction tuning alone, or its combination with pretraining, was responsible for eliciting L1 influence. Finally, we limited our L2 analysis to English. Extending this work to other, less data-rich L2s could prove insightful in future research.

Acknowledgements

This work was supported by funding from Cambridge University Press & Assessment. Compute facilities (32× NVIDIA A100-SXM4 GPUs) were provided by EleutherAI for model pretraining. With thanks to Krystian Balioskorski, Burak Bugrul, Aisha Delair, Ahmed Al-Jabri, Aidan Jones, Clare O'Driscoll, Rose Rociola, and EJ Zhou for their help with the translation of the survey. We also thank Bianca Ganescu, Przemek Kubiak, and Filip Trhlik for their work evaluating model outputs, and Dmytro Mai for technical assistance. We are grateful to Shiva Taslimipoor and Denise Löfflad for their helpful comments on an earlier draft of the manuscript. We thank the anonymous reviewers for their thoughtful comments.

Authors' Contributions. Conceptualisation: L.B.; Investigation: L.B., L.G., A.O.D.; Methodology: L.B., L.G., A.O.D.; Data curation: L.G.; Formal analysis: L.B., A.O.D.; Software: S.S. (pretraining), L.B., L.G., A.O.D. (instruction tuning and NLI); Funding acquisition: P.B., A.C.; Supervision: S.S., P.B., A.C., C.A.; Visualisation: L.G., L.B., A.O.D.; Resources: C.A., P.B., A.C.; Writing – original draft: L.B., L.G., A.O.D.; Writing – review and editing: L.B., L.G., A.O.D., S.S., C.A., A.C., P.B.

References

Tatsuya Aoyama and Nathan Schneider. 2024. Modeling nonnative sentence processing with L2 language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4927–4940, Miami, Florida, USA. Association for Computational Linguistics.

Catherine Arnett, Tyler A. Chang, James A. Michaelov, and Ben Bergen. 2025. On the acquisition of shared grammatical representations in bilingual language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20707–20726, Vienna, Austria. Association for Computational Linguistics.

Sarah Bernolet, Robert J Hartsuiker, and Martin J Pickering. 2007. Shared syntactic representations in bilinguals: Evidence for the role of word-order repetition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(5):931.

Barbara Berti, Andrea Esuli, and Fabrizio Sebastiani. 2023. Unravelling interlanguage facts via explainable machine learning. *Digital Scholarship in the Humanities*, 38(3):953–977.

BERTIN Project. 2023. alpaca-spanish: Spanish translation of the cleaned Stanford Alpaca dataset.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Zhenguang Cai, Xufeng Duan, David Haslett, Shuqi Wang, and Martin Pickering. 2024. Do large language models resemble humans in language use? In *Proceedings of the workshop on cognitive modeling and computational linguistics*, pages 37–56.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024. Monolingual or multilingual instruction tuning: Which makes a better Alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*.

Hasna Chouikhi, Manel Aloui, Cyrine Ben Hammou, Ghaith Chaabane, Haithem Kchaou, and Chehir Dhaouadi. 2024. Llamar & gemmar: Enhancing llms through arabic instruction-tuning.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Christine Cuskley, Rebecca Woods, and Molly Flaherty. 2024. The limitations of large language models for understanding human language and cognition. *Open Mind*, 8:1058–1083.

Mahdi Dhaini, Juraj Vladika, Ege Erdogan, Zineb Attaoui, and Gjergji Kasneci. 2025. Can LLM-generated textual explanations enhance model classification performance? an empirical study. In *International Conference on Artificial Neural Networks and Machine Learning (ICANN)*, pages 192–204. Springer.

Emplocity. 2023. Owca: Optimized and well-translated customization of alpaca. `https://huggingface.co/datasets/emplocity/owca`. Polish translation of the Alpaca instruction dataset. Accessed: 2026-03.

Rena Gao, Xuetong Wu, Tatsuki Kuribayashi, Mingrui Ye, Siya Qi, Carsten Roever, Yuanxing Liu, Zheng Yuan, and Jey Han Lau. 2025. Can llms simulate l2-english dialogue? an information-theoretic analysis of l1-dependent biases. *arXiv preprint arXiv:2502.14507*.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2013. Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project*, pages 240–254.

Paul Gonzalez-Torres and María del Pilar García Mayo. 2025. Grammatical transfer errors in EFL writing: Impact of Spanish influence, proficiency levels, and task types. *International Journal of Learning, Teaching and Educational Research*, 24(5):360–375.

Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie CY Chan, Andrew Lampinen, Jane X Wang, Zeynep Akata, and Eric Schulz. 2023. Machine psychology. *arXiv preprint arXiv:2303.13988*.

Robert J Hartsuiker, Martin J Pickering, and Eline Veltkamp. 2004. Is syntax separate or shared between languages? cross-linguistic syntactic priming in spanish-english bilinguals. *Psychological science*, 15(6):409–414.

Scott Jarvis. 2000. Methodological rigor in the study of transfer: Identifying l1 influence in the interlanguage lexicon. *Language learning*, 50(2):245–309.

Scott Jarvis and Scott A Crossley. 2012. *Approaching language transfer through text classification: Explorations in the detection-based approach*, volume 64. Multilingual Matters.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2026. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Transactions of the Association for Computational Linguistics*, 14:193–216.

Eojin Kang and Juae Kim. 2025. When language shapes thought: Cross-lingual transfer of factual knowledge in question answering. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 4868–4873.

Kristopher Kyle, Scott A Crossley, and You Jin Kim. 2015. Native language identification and writing proficiency. *International Journal of Learner Corpus Research*, 1(2):187–209.

Timothy P Lillicrap, Adam Santoro, Luke Marris, Colin J Akerman, and Geoffrey Hinton. 2020. Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(6):335–346.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, and 2 others. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. Fineweb-edu: the finest collection of educational content. Hugging Face.

Xiaoliang Luo, Michael Ramscar, and Bradley C Love. 2024. Beyond human-like processing: Large language models perform equivalently on forward and backward scientific text. *arXiv preprint arXiv:2411.11061*.

Ana Cristina Lahuerta Martínez. 2018. Analysis of syntactic complexity in secondary education efl writers at different proficiency levels. *Assessing Writing*, 35:1–11.

Yee Man Ng and Ilia Markov. 2025. Leveraging open-source large language models for native language identification. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 20–28, Abu Dhabi, UAE. Association for Computational Linguistics.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

Diane Nicholls, Andrew Caines, and Paula Buttery. 2024. The Write & Improve Corpus 2024: Error-annotated and CEFR-labelled essays by learners of English. Technical report. Cambridge University Press & Assessment.

Robin Nicholls and Kenneth Alperin. 2025. Cross-genre native language identification with open-source large language models. In *Proceedings of the 2nd LUHME Workshop*, pages 103–108, Bologna, Italy. UP - Universidade do Porto (https://doi.org/10.21747/978-989-9193-73-4/lan2), LIACC - Laboratório de Inteligência Artificial e Ciência de Computadores da Universidade do Porto, CLUP - Centro de Linguística da Universidade do Porto, UEF - The University of Eastern Finland and UAH - Universidad de Alcalá.

Jonathan Pacifico. 2024. French-Alpaca-dataset-Instruct-110K. 110k French instruction-response pairs generated with GPT-3.5-turbo in Alpaca format.

Suchir Salhan, Andrew Caines, and Paula Buttery. 2025. Pedagogical alignment of llms requires diverse cognitively-inspired student proxies. In *Proceedings of the First Workshop on CogInterp: Interpreting Cognition in Deep Learning Models at NeurIPS 2025*, San Diego, California, USA. NeurIPS Foundation.

Suchir Salhan, Ej Zhou, Laura Barbenel, Aoife O'Driscoll, Lily Goulder, Lucas Resck, Catherine Arnett, and Paula Buttery. 2026. Glints of gold or troubling waters? can a school of merged monolingual goldfish models swim in bilingual seas?

Larry Selinker. 1969. Language transfer. *General linguistics*, 9(2):67.

Itamar Shatz. 2017. Native language influence during second language acquisition: A large-scale learner corpus analysis. In *Proceedings of the Pacific Second Language Research Forum (PacSLRF 2016)*, pages 175–180. Japan Second Language Association Hiroshima, Japan.

Silk Road Project. 2023. Alpaca-data-gpt4-chinese. Chinese instruction-following dataset in Alpaca format (52k samples) with Chinese and English fields.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and 14 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Vivek Srivastava and Mayank Singh. 2021. Challenges and considerations with code-mixed nlp for multilingual societies. *Preprint*, arXiv:2106.07823.

Michael Swan and Bernard Smith. 2001. *Learner English: A teacher's guide to interference and other problems*, volume 1. Cambridge University Press.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the First Native Language Identification Shared Task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–57, Atlanta, Georgia. Association for Computational Linguistics.

TFLai. 2023. Turkish-alpaca. Turkish instruction-following dataset in Alpaca format (51.9k samples).

Ahmet Yavuz Uluslu, Tannon Kew, Tilia Ellendorff, Gerold Schneider, and Rico Sennrich. 2025. Robust native language identification through agentic decomposition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8398–8414, Suzhou, China. Association for Computational Linguistics.

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. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Sara Cushing Weigle. 2010. *Assessing Writing*. Cambridge University Press.

Ethan Gotlieb Wilcox, Michael Y Hu, Aaron Mueller, Alex Warstadt, Leshem Choshen, Chengxu Zhuang, Adina Williams, Ryan Cotterell, and Tal Linzen. 2025. Bigger is not always better: The importance of human-scale language modeling for psycholinguistics. *Journal of Memory and Language*, 144:104650.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. A survey on multilingual large language models: Corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362.

Aditya Yadavalli, Alekhya Yadavalli, and Vera Tobin. 2023. SLABERT talk pretty one day: Modeling second language acquisition with BERT. In *Proceedings of the 61st Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 11763–11777, Toronto, Canada. Association for Computational Linguistics.

Haiyin Yang, Zoey Liu, and Stefanie Wulff. 2025. Using NLI to identify potential collocation transfer in L2 English. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 687–696, Vienna, Austria. Association for Computational Linguistics.

Shuzhou Yuan, Zhan Qu, Mario Tawfelis, and Michael Färber. 2025. From monolingual to bilingual: Investigating language conditioning in large language models for psycholinguistic tasks. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1028–1040, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Wei Zhang and Alexandre Salle. 2023. Native language identification with large language models. *arXiv preprint arXiv:2312.07819*.

Yuwen Zhou and Yevgen Matusevych. 2025. Curse of bilinguality: Evaluating monolingual and bilingual language models on Chinese linguistic benchmarks. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 622–630, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Gustavo Zomer and Ana Frankenberg-Garcia. 2021. Beyond grammatical error correction: Improving L1-influenced research writing in English using pre-trained encoder-decoder models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2534–2540, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Methodology

A.1 L2LM Training

Training Hyperparameters Training hyperparameters are detailed in Table 6.

Parameter	Value / Notes
Tokenization	
Tokenizer type	Byte-Level BPE
Vocabulary size	50,000
Training corpus	L1 (CulturaX) + EN FineWeb-Edu datasets
Model Architecture	
Model type	GPT2LMHeadModel
Sequence length	512
Embedding dimension	896
Number of layers	24
Number of attention heads	14
BOS / EOS / PAD IDs	0 / 1 / 2
Training	
Batch size	16
Gradient accumulation	8 (effective batch size = 128)
Optimizer	AdamW (betas=(0.9,0.95), weight decay=0.1)
Learning rate	2e-4 (linear warmup + cosine decay)
Total training tokens	5B
Mixed precision	bfloat16 (CUDA autocast)
Gradient clipping	1.0 (norm)
GPUs / Distributed	Supports DDP, auto-detect rank/local rank
Checkpoint frequency	Curriculum-based (end of phase 1 + phase 2 L2 exposure stages: 25%, 50%, 75%, 100%)

Table 6: Hyperparameters for tokenization and model pretraining

Pretraining Each L2LM is pretrained on 8 NVIDIA A100-SXM4 GPUs (80GB of HBM2e memory per GPU), taking approximately 5 hours per model (Table 7). Preprocessing (data streaming and pretokenization) are CPU-bound, taking approximately 2 hours per model on a single node with 128 CPU threads.

Phase	Tokens (B)	Pretokenization (h)	Training Time (h)	Compute (GPU·h) per model
L1 dataset preprocessing	3.33	≈ 0.5	–	4 (8 GPUs × 0.5h)
EN dataset preprocessing	1.67	$\approx 0.25 - 0.3$	–	2.4 (8 GPUs × 0.3h)
Phase 1 training	2.5	–	2.5	20(8 GPUs × 2.5 h)
Phase 2 training	2.5	–	2.5	20(8 GPUs × 2.5 h)

Table 7: L2LM pretraining time per model.

Checkpoint	L1	L2	Total
Beginner (25%)	2.71B	0.42B	3.13B
Intermediate (50%)	2.92B	0.84B	3.75B
Advanced (75%)	3.13B	1.25B	4.38B

Table 8: Number of L1 and L2 tokens seen at each pretraining checkpoint.

A.2 Instruction Tuning

English-only dataset We used the Alpaca English dataset (Taori et al., 2023). It was filtered to remove any rows with an additional input (i.e., not just QA). After cleaning the dataset (e.g., removing empty rows), the dataset contained a total of 31,311 QA pairs.

English + L1 datasets We used the following Alpaca datasets for Spanish (BERTIN Project, 2023), French (Pacifico, 2024), German (Chen et al., 2024), Polish (Emplocity, 2023), Turkish (TFLai, 2023), Arabic (Chouikhi et al., 2024), and Chinese (Silk Road Project, 2023). After cleaning the datasets (e.g., removing non-text entries), we randomly selected 15,661 examples from each of the English and L1 datasets, yielding 31,322 examples in total. The datasets were normalised to ensure compatibility across different scripts.

Human L2 English instruction-tuning datasets These datasets were created by compiling essays from the W&I (Nicholls et al., 2024) and EFCAMDAT corpora (Geertzen et al., 2013). We included 10,000 QA pairs for all L1s except Polish (n = 370) and Turkish (n = 4,847) due to limited corpus data. All available essays from the W&I corpus were used, which is significantly smaller than the EFCAMDAT corpus. The EFCAMDAT contains relatively few C-level essays; all of these were included for each L1, with the remaining essays split evenly between A- and B-level according to corpus availability. This is illustrated in Figure 5.

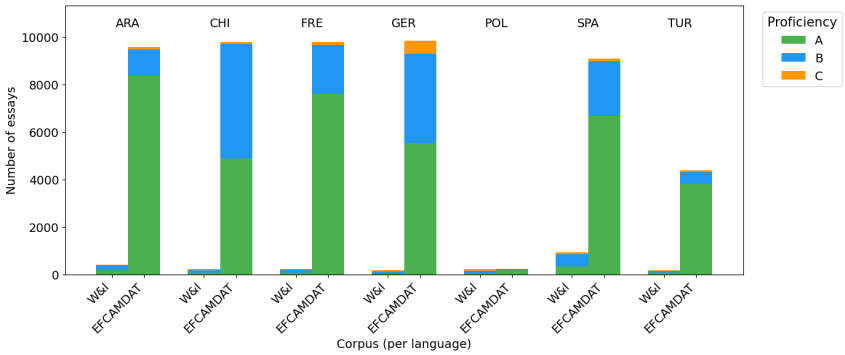


Figure 5: Proportion of essays in the L2 English instruction-tuning dataset from each corpus (W&I and EFCAMDAT) and proficiency level (A, B, C).

A.3 Prompt engineering

Text-generation prompting Text was generated using sampling (temperature = 0.7, top-p = 0.9) with a repetition penalty of 1.2 and blocked 4-grams. Outputs were constrained to 250-300 tokens (≈800-1,200 characters).

Prompt design Prompt design followed guidelines for low-proficiency learners (Weigle, 2010) to ensure accessibility across all three proficiency levels. The prompts given to L2 learners and LMs (with example L1 Spanish) are shown in Figure 6.

Prompt topic: Hobbies
Prompt style: Basic
Prompt:
Your town wants people to take part in different activities and hobbies. Write an article for your local newspaper describing the activities you can do in your town at the weekend.
Write around 150 words.

Prompt style: Mixed language
Prompt:
Your town wants people to take part en diversas actividades y aficiones. Write an article para tu periódico local describing the activities you can do in your town durante el fin de semana.
Write around 150 words.

Prompt style: L2T
Prompt:
Your town wants people to take part in different activities and hobbies. Write an article for your local newspaper describing the activities you can do in your town at the weekend.
Think in Spanish. Write around 150 words in English.

(a) Hobbies.

Prompt topic: Family
Prompt style: Basic
Prompt:
Your local magazine is holding a competition for articles about the question 'Big family or small family'. Write an article about your family. You should include whether you have a big or small family, and which you prefer.
Write around 150 words.

Prompt topic: Family
Prompt style: Mixed language
Prompt:
Tu revista local is holding a competition for articles about the question 'Big family or small family'. Escribe un artículo sobre tu familia. You should include whether you have a big or small family, y cual prefieres.
Write around 150 words.

Prompt topic: Family
Prompt style: L2T
Prompt:
Your local magazine is holding a competition for articles about the question 'Big family or small family'. Write an article about your family. You should include whether you have a big or small family, and which you prefer.
Think in Spanish. Write around 150 words in English.

(b) Family.

Prompt topic: Holidays
Prompt style: Basic
Prompt:
Your friend, Sophia, has emailed you about her recent holiday. Write an email to her, describing your favourite holiday you have been on.
Write around 150 words.

Prompt style: Mixed language
Prompt:
Tu amiga, Sophia, has emailed you about her recent holiday. Write an email to her, describiéndole tus vacaciones favoritas.
Write around 150 words.

Prompt style: L2T
Prompt:
Your friend, Sophia, has emailed you about her recent holiday. Write an email to her, describing your favourite holiday you have been on.
Think in Spanish. Write around 150 words in English.

(c) Holiday.

Figure 6: The nine generation prompts.

A.4 Evaluation

Named Entity Recognition The following named entities were considered to potentially encode L1-identifying information: location (LOC), Geopolitical Entity (GPE), PERSON and LANGUAGE. They were replaced with following entities thought to be English-related and thus L1-neutral: England (LOC), London (GPE), Alex (PERSON) and English (LANGUAGE).

Manual edits for human evaluation For the human evaluation, one evaluator was given L2LM essays after editing to remove generation artefacts. We adopted a conservative approach, retaining all full sentences regardless of semantic content or fluency. We removed only formatting abnormalities including links, symbols, bracketed text and repetition of single words.

Before Editing	We must always be confident in your opinion and be sure to make it possible. The news will likely change too quickly as we know what happened before and why. You can also be thankful for the best advice and support from our readers. Taking care of yourself and helping others with difficult issues could lead to more positive conversations than ever before. By using these tips, you can be an effective reader and a great source of information. Thank you! Sincerely,[Your Name] [Company Name] [Name], [Comporting Name] (https://www.newadventureofthewineandlife.com/imgurst-topic) [Compromising Name] [Description] [Images/] <nooutput type=""submit"" style=""text"" name=""text"">This article offers a range of helpful facts and evidence that may not be available online or offline. For example, if you're looking for a book or book, you would need to use it to get your headphones on and do some research. Your thoughts and suggestions might not necessarily reflect your opinions or feelings, but they may not reflect your beliefs and ideas. If you are discussing this topic online, please don't hesitate to contact us at least immediately. Thank you for your time and consideration.Sincerely. [Insert Name] Sincerately,[Your Email Us
After Editing	We must always be confident in your opinion and be sure to make it possible. The news will likely change too quickly as we know what happened before and why. You can also be thankful for the best advice and support from our readers. Taking care of yourself and helping others with difficult issues could lead to more positive conversations than ever before. By using these tips, you can be an effective reader and a great source of information. This article offers a range of helpful facts and evidence that may not be available online or offline. For example, if you're looking for a book or book, you would need to use it to get your headphones on and do some research. Your thoughts and suggestions might not necessarily reflect your opinions or feelings, but they may not reflect your beliefs and ideas. If you are discussing this topic online, please don't hesitate to contact us at least immediately. Thank you for your time and consideration.Sincerely.

Figure 7: Generation from a Spanish beginner model instruction tuned on Alpaca English answering the mixed prompt with family topic before and after editing.

B Results: RQ1 additional materials

For the NLI classification task, we used the following system and user prompts (Figure 8):

```
You are a forensic linguistics expert that reads English texts written by
non-native authors to classify the native language of the author as one
of:

"SPA": SPANISH
"FRE": FRENCH
"CHI": CHINESE
"GER": GERMAN
"POL": POLISH
"ARA": ARABIC
"TUR": TURKISH

Use clues such as spelling errors, word choice, syntactic patterns, and
grammatical errors to decide on the native language of the author.

DO NOT USE ANY OTHER CLASS.
IMPORTANT: Do not classify any input as "ENG" (ENGLISH). English is an
invalid choice.

You must provide a guess. Output two named sections: (1) "Class" with the
name of the language, and (2) "Reasoning" with an explanation of your
judgement with examples from the text.

You must respond with a single valid JSON object in the specified
format.

USER_PROMPT =

<Input Text>
{text}

Classify the text as ONLY ONE of: SPA, FRE, CHI, GER, POL, ARA, TUR. Do
not output any other class- do NOT choose "ENG" (ENGLISH). What is the
closest native language of the author of this English text from the given
list?

Respond in the following JSON format:

"Class": "One of SPA, FRE, CHI, GER, POL, ARA, TUR",
"Reasoning": "An explanation of your judgement with examples from the
text"
```

Figure 8: System and user prompts for NLI.

B.1 Experiment 1

Differences in NLI accuracy across the five runs were not significant for any model condition (Table 9a), indicating that performance was consistent and reproducible. Accuracy did not differ significantly across L1s for any model condition (Table 9b).

	Q(4)	p
GPT-4 explanation	6.40	.17
GPT-4 no explanation	2.11	.72
GPT-5 explanation	4.25	.37
GPT-5 no explanation	1.56	.82

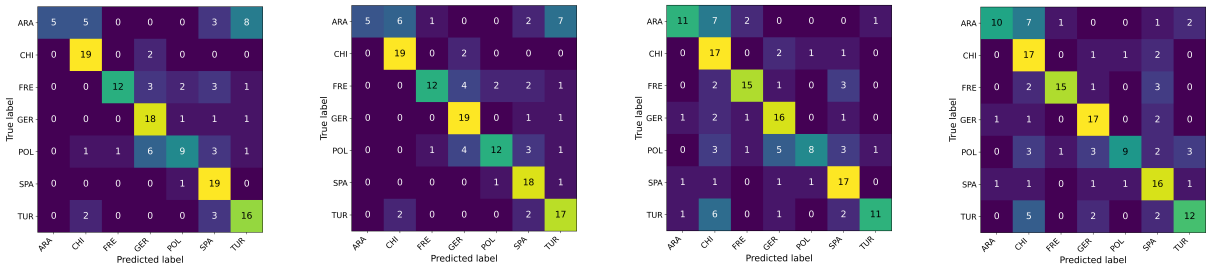
(a) Cochran's Q tests across five iterations for each model condition.

	$\chi^2(6)$	p
GPT-4 explanation	10.40	.11
GPT-4 no explanation	8.85	.18
GPT-5 explanation	4.98	.55
GPT-5 no explanation	5.58	.47

(b) Kruskal-Wallis rank sum test results for model accuracy across L1s.

Table 9: Statistical test results for NLI model accuracy.

Confusion matrices of the NLI predictions for each model condition on the W&I subset (21 essays per L1) are shown in Figure 9.



(a) GPT-4 (explanation). (b) GPT-4 (no explanation). (c) GPT-5 (explanation). (d) GPT-5 (no explanation).

Figure 9: Confusion matrices for all model conditions.

The omission of articles ('I used to work on the computer aproximetly 3 to 4 hours')

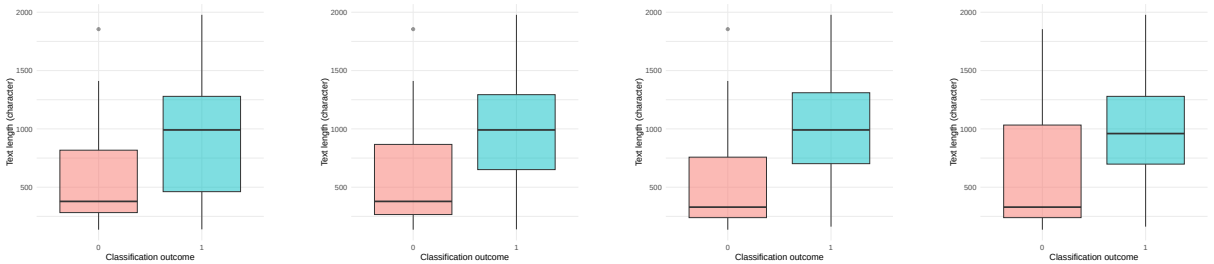
“celebrate your birthday” is a common calque from “celebrar” (more natural in English would be “celebrate your birthday” is possible, but the overall phrasing is Spanish-like)

The use of 'pannel' instead of 'panel' is likely influenced by the French 'panel.'

Figure 10: Examples of hallucinations/inaccuracies generated by GPT-4 in the explanation-prompting condition.

B.2 Experiment 2

Original W&I subset Figure 11 and Table 10 show that NLI classification outcomes (correct vs. incorrect) are significantly affected by essay length (in characters) across all four model conditions. Correctly classified essays are significantly longer than incorrectly classified essays. As shown in Table 11, NLI accuracy on A-level essays is significantly lower than for C-level essays across all four model conditions.



(a) GPT-4 (explanation). (b) GPT-4 (no explanation). (c) GPT-5 (explanation). (d) GPT-5 (no explanation).

Figure 11: Boxplots of text length by classification accuracy (averaged over five runs). Texts with mean accuracy <1 are classified as 0; mean accuracy = 1 are classified as 1.

Model condition	Incorrect mean	Correct mean	t(df)	p
GPT-4 (explanation)	551.7	955.3	-4.23(40)	<.001
GPT-4 (no explanation)	591.9	965.2	-4.1(53)	<.001
GPT-5 (explanation)	535.2	988.9	-5.1(57)	<.001
GPT-5 (no explanation)	562.4	980.7	-4.5(54)	<.001

Table 10: Classification outcome (accuracy averaged over five runs) by mean text length (characters); instances with mean accuracy = 1 are labelled as correct, otherwise incorrect. Statistics are from Welch two-sample t-tests.

Model condition	Comparison (Direction)	W	p
GPT-4 (explanation)	A vs B (A < B)	1038	.39
GPT-4 (explanation)	A vs C (A < C)	834	<.001
GPT-4 (explanation)	B vs C (B < C)	1004	<.05
GPT-4 (no explanation)	A vs B (A < B)	977	.14
GPT-4 (no explanation)	A vs C (A < C)	847	<.01
GPT-4 (no explanation)	B vs C (B < C)	1080	.44
GPT-5 (explanation)	A vs B (A < B)	818	<.01
GPT-5 (explanation)	A vs C (A < C)	740	<.001
GPT-5 (explanation)	B vs C (B < C)	1106	.72
GPT-5 (no explanation)	A vs B (A < B)	796	<.001
GPT-5 (no explanation)	A vs C (A < C)	785	<.001
GPT-5 (no explanation)	B vs C (B < C)	1196	1.0

Table 11: Post-hoc Bonferroni-corrected pairwise Wilcoxon tests by model condition on non-truncated essays.

Truncated W&I subset A Friedman test found that there was no significant difference in accuracy between conditions, $\chi^2(3) = 1.08, p = .78$. As summarised in Table 12, NLI accuracy did not differ significantly between model conditions for each proficiency level, nor between proficiency level within each model condition.

	Overall	A	B	C
GPT-4 explanation	67	67	69	65
GPT-4 no explanation	67	64	70	67
GPT-5 explanation	65.6	55	70	72
GPT-5 no explanation	66	54	70	74

(a) NLI accuracy (%) by model and proficiency (truncated essays).

	$\chi^2(2)$	p
GPT-4 explanation	0.09	.96
GPT-4 no explanation	0.45	.80
GPT-5 explanation	3.70	.16
GPT-5 no explanation	5.25	.07

(b) Kruskal–Wallis tests across proficiency (A/B/C).

	$\chi^2(3)$	p
A proficiency	6.21	.10
B proficiency	0.27	.97
C proficiency	2.26	.52

(c) Friedman tests across models within each proficiency.

Table 12: Summary of NLI results for truncated essays.

B.3 Novel L2 learner dataset

Participants were asked to self-report their proficiency using the CEFR scale; guidelines to proficiency levels (both in English and the L1) were provided as a reference. Participants were compensated at a rate equivalent to the U.K.'s 'real living wage' set by the Living Wage Foundation[5].

Each participant responded to three prompts, each with a distinct style and topic. Due to technical issues, 31 participants were unable to respond to all three prompts; we included the responses that were available. One participant's responses were excluded due to non-compliance with the task, as they consisted of only a few sentences and excessive punctuation to meet the minimum character limit.

Essays were manually screened for signs of generative AI usage; however, no essays were excluded on this basis.

The dataset comprises 466 essays with a mean length of 906 (SD = 109) characters. A per-L1 breakdown of our dataset is provided in Table 13.

[5]At a non-London rate of £13.45 at the time of collecting data in 2026; https://www.livingwage.org.uk/what-real-living-wage

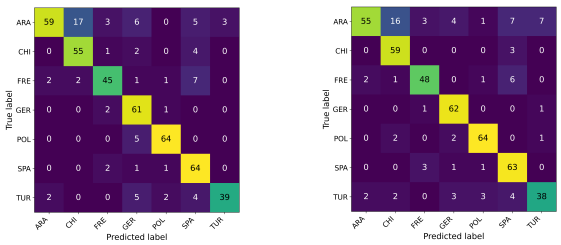
L1	no. essays	no. participants
Spanish	68	25
French	58	20
German	64	22
Polish	69	28
Turkish	52	18
Arabic	93	40
Chinese	62	21

Table 13: Composition of our L2 learner English dataset.

A GLMM was fitted with L1, proficiency, prompt style, prompt topic, and NER condition as fixed effects, and user ID as a random intercept. Model summary is provided in Table 14. Whilst accuracy for L1 Turkish and Arabic essays was significantly different from the intercept, a binomial test confirmed that these essays were classified significantly above chance. There was no significant effect of prompt style or topic. NER normalisation did not affect NLI accuracy (see Figure 12). Crucially, proficiency had no significant effect on NLI accuracy.

Term	Estimate	Std. Error	z value	p
(Intercept)	5.15	1.04	4.96	<.001
L1: French	-1.72	1.05	-1.64	.10
L1: German	1.01	1.20	0.84	.40
L1: Polish	-0.41	1.02	-0.40	.69
L1: Turkish	-3.12	1.07	-2.92	<.01
L1: Arabic	-3.99	0.96	-4.14	<.001
L1: Chinese	-0.21	1.08	-0.20	.84
NER	0.06	0.24	0.24	.81
Proficiency: Intermediate	-0.09	0.69	-0.13	.89
Proficiency: Advanced	-1.16	0.70	-1.67	.09
Style: L2T	0.44	0.31	1.42	.16
Style: Mixed	0.09	0.30	0.30	.77
Topic: Hobbies	-0.33	0.30	-1.11	.27
Topic: Holiday	0.09	0.31	0.30	.76

Table 14: Model summary for NLI accuracy on the L2 learner dataset. The intercept is set to Spanish, basic prompt style, family topic, beginner proficiency, no NER.

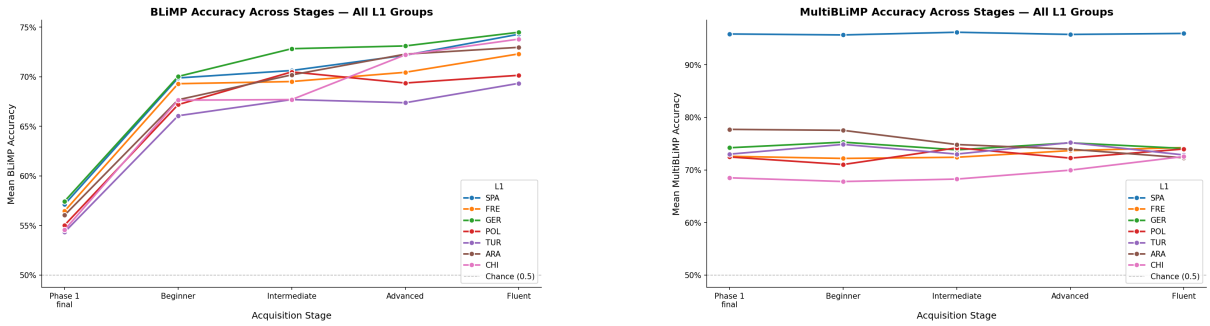


(a) Original (b) NER-normalised

Figure 12: Confusion matrices for NLI accuracy on our novel L2 dataset: original and NER-normalised essays.

C Results: RQ2 results additional materials

C.1 BLiMP and MultiBLiMP



(a) BLiMP results

(b) MultiBLiMP results

Figure 13: BLiMP and MultiBLiMP accuracy for the seven base models.

Training stage	BLiMP	MultiBLiMP
Phase 1	.85	.42
Beginner	.86	.42
Intermediate	.95	.42
Advanced	.76	.42
Fluent	.73	.42

Table 15: Kruskal-Wallis test of whether L1 was a significant predictor of BLiMP and MultiBLiMP accuracy. p-values are reported.

C.2 Comparing instruction-tuning conditions

Term	Estimate	Std. Error	z value	p
(Intercept)	-2.92	0.48	-6.07	<.001
L1: French	-0.38	0.31	-1.21	.23
L1: German	-0.22	0.31	-0.70	.49
L1: Polish	-3.15	0.41	-7.73	<.001
L1: Turkish	-1.63	0.35	-4.66	<.001
L1: Arabic	-1.70	0.34	-4.93	<.001
L1: Chinese	5.25	0.44	11.81	<.001
IT: English-only	-0.16	0.52	-0.31	.76
IT: English + L1	0.52	0.48	1.08	.28
IT: L2 English	6.75	0.51	13.33	<.001
NER	0.03	0.26	0.13	.90
Proficiency: Intermediate	-0.04	0.24	-0.17	.86
Proficiency: Advanced	0.12	0.23	0.51	.61
Style: L2T	0.10	0.48	0.22	.83
Style: Mixed	2.75	0.43	6.43	<.001
Topic: Hobbies	0.64	0.31	2.04	<.05
Topic: Holiday	0.69	0.31	2.19	<.05
IT: English-only * NER	0.04	0.38	0.12	.91
IT: English + L1 * NER	-0.12	0.35	-0.34	.73
IT: L2 English * NER	-3.37	0.40	-8.35	<.001
Style: L2T * Topic: Hobbies	0.25	0.41	0.60	.55
Style: L2T * Topic: Holiday	-0.09	0.41	-0.22	.83
Style: Mixed * Topic: Hobbies	-1.03	0.39	-2.62	<.01
Style: Mixed * Topic: Holiday	-0.58	0.39	-1.50	.13
IT: English-only * Style: L2T	1.34	0.56	2.41	<.05
IT: English-only * Style: Mixed	-1.22	0.53	-2.33	<.05
IT: English + L1 * Style: L2T	0.69	0.52	1.31	.19
IT: English + L1 * Style: Mixed	-0.70	0.47	-1.48	.14
IT: L2 English * Style: L2T	-0.08	0.49	-0.16	.87
IT: L2 English * Style: Mixed	-2.22	0.44	-5.00	<.001

Table 16: Model summary for NLI accuracy across instruction-tuning conditions. The intercept is set to Spanish L1, base model, basic prompt style, family prompt topic, beginner proficiency, and original essay (no NER) condition.

Prompt	**Votre magazine local is holding a competition for articles about the question 'Big family or small family'. Écrivez un article sur votre famille. You should include whether you have a big or small family, et laquelle vous préférez. Write around 150 words.**
L2 English IT	it's my opinion that big family and small family are really important. As we all know, our children will be very friendly but not as lazy as us. We can say that it's difficult to talk with their parents because they have different opinions about them. Moreover, the baby of our sons is quite easy to make up all these views. This is a good idea because the young man is looking forward to being able to help his father when he wants. The younger one is looking forward and able to move in order to see if his father is successful in this position. It's more appropriate to keep your mind on what our society will look like before the next birthday. I hope you can come! Bye! Rene B.C.E.F.I sincerely, Your message : Hi! Our family is beautiful but not so smart. My best friend, Anna, is confident and hard-working. She has many friends who don't understand me and she never think about it, she always talks on the phone. She doesn't feel like her mother. She doesnt look at me and she's afraid of the future. He is fairly aggressive. If you follow me, let me know if you're right. See you soon!
Base	A person who has a big family or small family has not yet had a big family. Your magazine local is holding a competition for articles about the question 'Big family or small family'. Your magazine local is holding a competition for articles about the question 'Big family or small family'. Votre magazine local is holding a competition for articles about the question 'Big family or small family'. Your magazine local is holding a competition for articles about the question 'Big family or small family'. Your magazine local is holding a competition for articles about the question 'Big family or small family'. Your magazine local is holding a competition for articles about the question 'Big family or small family'. You should include whether you have a big family and your family. You should cite the title of the journal, which is usually the title of the journal.### Response:bilaterally ### The ### Questions about the Question ### Answers to the ### Questions about the Question ### Questions about the Question ### Questions about the Question ### Questions about the Question ### Questions about the Question ### Questions about the Question ### Questions about the Question ### Questions about the Question ### Questions about the Question ### Questions about the Question ### Questions about the Question ### Questions about the Question ### Questions about the Question ### Questions about the Question ### Questions about the Question ### Questions about the Question ### Questions about the Question ### Questions about the Question ### Questions about the Question

Figure 14: A comparison of the generated essays from French intermediate model with L2 English IT vs. base model. This is the mixed-language prompt style and the family prompt topic.

C.3 Base model

378 essays were generated (54 per L1), with a mean length of 1,204 (SD = 316) characters. Mixed-language prompt style significantly improved NLI accuracy (see Table 17). This was robust even when Chinese as L1 was removed (i.e., the effect was not driven by Chinese false positives). NLI accuracy on mixed-language prompted essays was significantly above chance ($p < .001$); whilst this could be taken as an indication of the production of L1-influenced L2 English, a manual inspection of the 47 correctly predicted essays concluded that 24 of these (51%) were written at least partly in the L1, and 16 were written mostly or fully in the L1.

Term	Estimate	Std. Error	z value	p
(Intercept)	-2.81	0.56	-5.06	<.001
L1: French	0.59	0.55	1.07	.28
L1: German	0.31	0.56	0.56	.58
L1: Polish	-1.64	0.84	-1.94	.05
L1: Turkish	-2.37	1.10	-2.15	<.05
L1: Arabic	-0.18	0.60	-0.30	.77
L1: Chinese	4.71	0.67	7.01	<.001
Style: L2T	-0.00	0.53	0.00	1.0
Style: Mixed	2.10	0.47	4.47	<.001

Table 17: Model summary for NLI accuracy on the original base model texts. The fixed effects are L1 and prompt style. The intercept is set to Spanish L1 and basic prompt style.

In contrast, none of the 16 correctly predicted L1-Chinese essays contained Chinese words or characters, suggesting that NLI may rely on subtle cues or be influenced by an overprediction bias; further investigation is needed to determine the underlying reason. Excluding these L1-Chinese essays, 24/31 (77.4%) essays were written at least partly in the L1. Ultimately, NLI accuracy alone cannot be taken as an indicator of L1-influenced L2 text, and must be verified with manual inspection of the essays.

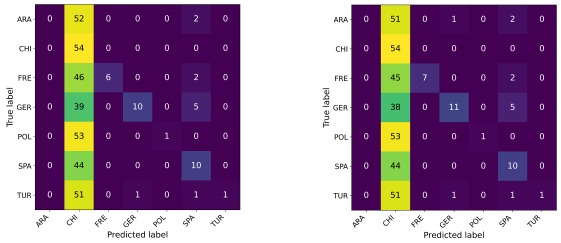
C.4 English-only instruction-tuning

A total of 378 essays (54 per L1) were generated with a mean length of 1,121 (SD = 197) characters. Overall NLI accuracy was 21.69% (95%CI[17.64, 26.19]). Although overall accuracy was significantly above chance ($p < .001$), this result was largely driven by perfect recall (1.0) for Chinese essays. As

shown in Figure 15, this appears to reflect a strong overprediction bias towards Chinese, with low precision (0.16) due to a high number of false positives. Indeed, per-L1 accuracy was not above chance for any L1s other than Chinese (Table 18).

L1	F1	P	R	*p*
Spanish	0.27	0.50	0.19	.24
French	0.20	1.0	0.11	.80
German	0.31	0.91	0.19	.24
Polish	0.36	1.0	0.02	1.0
Turkish	0.04	1.0	0.02	1.0
Arabic	0.00	0.00	0.00	1.0
Chinese	0.27	0.16	1.0	<.001

Table 18: Precision (P), Recall (R), and F1 scores per L1 background for the original (non-NER) essays from English-only instruction tuning. p-values are from a one-tailed binomial test.



(a) Original essays.

(b) NER-normalised essays.

Figure 15: Confusion matrices for NLI on essays generated with English-only instruction tuning: original vs. NER-normalised essays.

C.5 English and L1 instruction-tuning

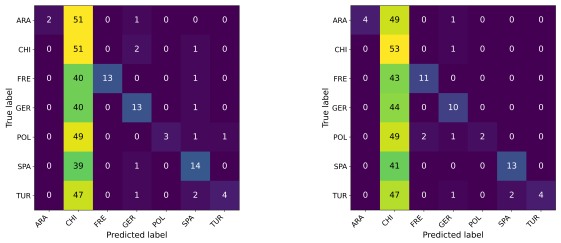
A total of 378 essays were generated (54 per L1) with a mean length of 1,169 (SD = 228) characters. Overall NLI accuracy was 26.46% (95%CI[22.08, 31.21]). Although overall accuracy was significantly above chance ($p < .001$), this result was largely driven by near-ceiling recall (0.94) for Chinese essays.

As shown in Figure 16, this appears to reflect a strong overprediction bias towards Chinese, with low precision (0.16) due to a high number of false positives. Per-L1 accuracy (Table 19) was above chance for Chinese and Spanish, and below chance for Arabic.

After NER normalisation (see Figure 16), NLI accuracy was significantly above chance for Chinese, and significantly below chance for Polish. A manual inspection of the 13 correctly predicted L1-Spanish essays revealed that six contained entire sentences in Spanish, with one essay fully written in Spanish.

L1	English + L1 (original)				English + L1 (NER)			
	F1	P	R	*p*	F1	P	R	*p*
Spanish	0.38	0.70	0.26	<.05	0.38	0.87	0.24	.05
French	0.39	1.0	0.24	.05	0.33	0.85	0.20	.24
German	0.36	0.72	0.24	.05	0.29	0.71	0.19	.34
Polish	0.11	1.0	0.06	.08	0.07	1.0	0.04	<.05
Turkish	0.14	0.80	0.07	.18	0.14	1.0	0.07	.18
Arabic	0.07	1.00	0.04	<.05	0.14	1.0	0.07	.18
Chinese	0.27	0.16	0.94	<.001	0.28	0.16	0.98	<.001

Table 19: Precision (P), Recall (R), and F1 scores per L1 background for English + L1 instruction tuning condition: original and NER-normalised essays. p-values are from a two-sided binomial test.



(a) Original essays.

(b) NER-normalised essays.

Figure 16: Confusion matrices for NLI on essays generated with English and L1 instruction tuning: original vs. NER-normalised essays.

C.6 L2 English instruction tuning

A total of 378 essays were generated (54 per L1) with a mean length of 1,013 ± 152 characters. To determine the effects of prompt style and prompt topic, a GLM was fitted to NLI accuracy on the original essays (i.e., without NER normalisation) in this instruction-tuning condition. The model included prompt style and prompt topic as fixed effects, as well as their interactions. The model summary is in Table 20. Large standard errors are present due to ceiling effects in the data. No effects were statistically significant.

Term	Estimate	Std. Error	z value	p
(Intercept)	2.00	0.48	4.20	<.001
Style: L2T	0.25	0.71	0.35	.73
Style: Mixed	-0.55	0.62	-0.90	.37
Topic: hobbies	-0.21	0.65	-0.32	.75
Topic: holiday	1.71	1.12	1.53	.13
Style: L2T * Topic: hobbies	17.52	1659.38	0.01	.99
Style: L2T * Topic: holiday	15.60	1659.38	0.01	.99
Style: Mixed * Topic: hobbies	1.01	0.92	1.10	.27
Style: Mixed * Topic: holiday	0.55	1.56	0.36	.72

Table 20: Model summary: L2 English IT condition, original version of essays. The intercept is set to basic prompt style and family prompt topic.

NER-normalised essays were classified significantly less accurately than the original ones for all L1s except Spanish. McNemar's test was used when possible. For L1s where original accuracy was 100% (Arabic and Chinese), a binomial test on the discordant pairs was used instead as McNemar's test cannot be computed (see Table 21).

L1	p-value
Spanish	.13
French	< .001
German	< .001
Polish	< .001
Turkish	< .001
Arabic	< .001
Chinese	< .001

Note. Binomial test used for Arabic and Chinese where original essay accuracy = 100%.

Table 21: NLI accuracy on L2 English IT: original vs. NER-normalised essays.

Prompt	**Your town wants people to take part in different activities and hobbies. Write an article for your local newspaper describing the activities you can do in your town at the weekend. Think in Polish. Write around 150 words in English.**
L2 English IT (Original)	In my town there is some fun, but it's a big thing. I'm quite tired because of my life in **Poland**.I'm very anxious to get the food we eat during the day.I have been getting food at home on Saturday afternoon. He doesn't like any other activity, but he'll always go with me every morning. You know how wonderful you are! My family has a lot of children. I love the way of talking. I think this kind of time can be enjoyable if you want to travel. My friends are looking forward and they're going to find something which makes them happy! I hope that I will stay at school or maybe even longer. Let's say - everyone likes going to work when you want to make your job. The first thing about learning English is to learn English. It's hard to say what you need to do when you want yourself. We don't need to come up with another person, especially if you want him/her. They will talk about things which helps us to move around their lives. That might help you feel more comfortable. So let's take care. I am sure we will meet you. There is plenty of resources available. My name is **Anna**. I really wanted to live in **Warsaw**.
L2 English IT (NER)	In my town there is some fun, but it's a big thing . I'm quite tired because of my life in **London** . I'm very anxious to get the food we eat during the day. I have been getting food at home on Saturday afternoon . He doesn't like any other activity , but he'll always go with me every morning. You know how wonderful you are! My family has a lot of children. I love the way of talking. I think this kind of time can be enjoyable if you want to travel. My friends are looking forward and they're going to find something which makes them happy! I hope that I will stay at school or maybe even longer. Let 's say - everyone likes going to work when you want to make your job. The first thing about learning English is to learn English. It 's hard to say what you need to do when you want yourself. We don't need to come up with another person, especially if you want him/her. They will talk about things which helps us to move around their lives. That might help you feel more comfortable. So let 's take care. I am sure we will meet you. There is plenty of resources available. My name is **Alex**. I really wanted to live in **London**.

Figure 17: A comparison of the generated essays from Polish beginner, L2 English IT model condition: original vs. NER-normalised essays. This is the L2T prompt style and the hobbies prompt topic.

C.7 Human vs. LM

Term	Estimate	Std. Error	z value	p
(Intercept)	4.13	0.76	5.45	<.001
L1: French	-1.54	0.79	-1.95	.051
L1: German	0.82	0.93	0.88	.38
L1: Polish	-0.33	0.79	-0.42	.67
L1: Turkish	-2.48	0.79	-3.12	<.01
L1: Arabic	-3.25	0.72	-4.54	<.001
L1: Chinese	-0.17	0.84	-0.20	.84
Animacy: LM	1.46	1.34	1.09	.28
NER	0.05	0.22	0.23	.82
Style: L2T	0.24	0.20	1.18	.24
Style: Mixed	0.04	0.20	0.22	.83
Topic: Hobbies	-0.30	0.28	-1.07	.28
Topic: Holiday	0.06	0.28	0.21	.83
Proficiency: Intermediate	0.01	0.46	0.01	.99
Proficiency: Advanced	-0.71	0.46	-1.55	.12
Animacy: LM * L1: French	-0.40	1.71	-0.23	.82
Animacy: LM * L1: German	-2.37	1.78	-1.33	.18
Animacy: LM * L1: Polish	-5.33	1.74	-3.07	<.01
Animacy: LM * L1: Turkish	0.60	1.71	0.35	.72
Animacy: LM * L1: Arabic	0.28	1.67	0.17	.87
Animacy: LM * L1: Chinese	3.21	2.12	1.51	.13
Animacy: LM * NER	-3.94	0.46	-8.66	<.001
Animacy: LM * Topic: Hobbies	0.89	0.40	2.24	<.05
Animacy: LM * Topic: Holiday	1.00	0.41	2.46	<.05

Table 22: Model summary. The intercept is set to human-written, Spanish L1, basic prompt style, family topic, and original essays (no NER).

C.8 Human evaluation

Evaluators were three times more likely to misclassify human-written texts as LM-generated than vice versa (Figure 18; however, a McNemar's test indicated that this difference was not statistically significant (p = .08). We tested the effects of essay L1, instruction-tuning condition, prompt style, prompt topic, proficiency, and whether NLI prediction was correct; none of these variables significantly affected

evaluation outcomes (Table 23).

Condition	p
L1	.56
Data condition	.13
Prompt Style	1.0
Prompt Topic	.42
Proficiency	.89
Correct NLI	.19

Table 23: Fisher's exact tests for significance in human evaluation.

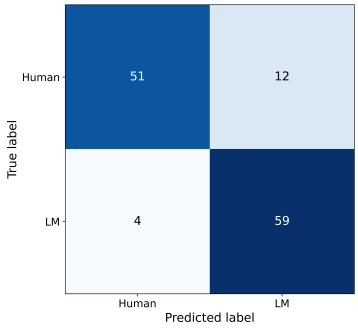


Figure 18: Confusion matrix of binary human evaluation classifying essays as human-written or LM-generated.

A Scalable Tool for Measuring Manner and Result Verbs in Developmental Language Research

Divyesh Pratap Singh
University at Buffalo

Dakshesh Gusain
Nanyang Technological University

Federica Bulgarelli
University at Buffalo

Alison Eisel Hendricks
University at Buffalo

John Beavers
The University of Texas at Austin

Nathan M. Beers
University at Buffalo

Ifeoma Nwogu
University at Buffalo
inwogu@buffalo.edu

Abstract

Manner and result verbs encode different aspects of event structure and have been discussed in developmental work as a potentially informative distinction for studying early verb learning. However, this distinction remains difficult to measure at scale because large annotated resources for manner and result classification are not currently available. We present a computational approach for identifying manner and result verbs in sentence context. Using linguistically informed prompts, we generate sentence-level annotations with large language models over data drawn from MASC and InterCorp, extending coverage from previously annotated portions of VerbNet to 436 classes. We then train a RoBERTa-based classifier on these annotations and evaluate it on three held-out gold-standard datasets, including previously annotated items and a new expert-annotated set. Across these evaluations, the model shows promising performance, with average accuracy up to 89.6%. We present this work as a scalable measurement tool that can support future research on verb semantics in developmental and other language datasets, while noting that further validation is needed for borderline cases, mixed manner/result verbs, and downstream developmental applications.

1 Introduction

Early language development depends not only on how much language children hear, but also on the kinds of meanings encoded in the words they learn. Verbs are especially important because they support children's transition to multiword speech and later grammatical development. Verb vocabulary around age two predicts later grammatical outcomes and, for some developmental questions, may be more informative than noun vocabulary (Hadley et al., 2016). These issues are especially relevant for late talkers, whose early language trajectories are heterogeneous and whose later outcomes are difficult to predict from broad lexical measures alone.

One semantic distinction that has become relevant in this literature is the contrast between *manner* and *result* verbs. Manner verbs encode how an action is carried out (e.g., *rub*, *scribble*, *run*), whereas result verbs encode an outcome or change of state (e.g., *clean*, *fill*, *open*) (Hovav and Levin, 2010; Levin, 2008). Developmental work suggests that this distinction may be informative for understanding variation in early verb learning. For example, Horvath et al. (2022) report that the relative proportions of manner and result verbs differ between late talkers and typically developing children, and that children who produce more manner verbs also tend to produce more verbs overall (Horvath et al., 2019, 2022). Because a substantial proportion of children with early language delay later meet the criteria for Developmental Language Disorder (DLD), the task of identifying finer-grained semantic properties of children's early vocabularies may help clarify which aspects of early language are associated with these later outcomes.

At the same time, this distinction remains difficult to study at scale. Although computational linguistics has made substantial progress on grammatical annotation tasks such as part-of-speech tagging (DeRose, 1988), fine-grained semantic categorization is generally more challenging. Prior work on related event-semantic distinctions suggests that verb meaning is difficult to classify automatically in context (Friedrich et al., 2022; Friedrich and Gateva, 2017; Metheniti et al., 2022; Friedrich et al., 2016). As a result, theoretically important contrasts such as manner versus result verbs still lack broad, scalable annotation resources, limiting their use in developmental language research.

To address this gap, we present a computational approach for identifying manner and result verbs in context. We use large language models (LLMs) as informed annotators, drawing on established linguistic definitions of manner and result verbs together with a small set of illustrative examples. We

Proceedings of the 1st Workshop on Computational Developmental Linguistics (CDL), pages 117–128

prompt LLMs to label sentences from the Manually Annotated Sub-Corpus (MASC; Ide et al. 2008) and the InterCorp parallel corpus (ek Čermák and Rosen 2012), expanding coverage from 151 previously annotated VerbNet classes to 436 classes (Brown et al., 2019; Kipper et al., 2008). We then fine-tune a pretrained RoBERTa classifier (Liu et al., 2019) on these labels and evaluate it on three held-out gold-standard datasets. We position this system as a scalable measurement tool that can support research on verb semantics in larger corpora, including developmental language data.
In summary, our contributions are:

- We present a scalable computational framework for identifying *manner* and *result* verbs in sentence context, enabling this theoretically important distinction to be measured in larger language datasets.
- We introduce an annotation pipeline that leverages large language models to generate training data for this task in the absence of large-scale gold-standard resources.
- We show that a RoBERTa-based classifier trained on these annotations can reliably distinguish manner and result verbs in context.
- We will publicly release our code and annotated dataset, extending coverage to 436 VerbNet classes, to support future research.

2 Understanding Verb Root Meaning

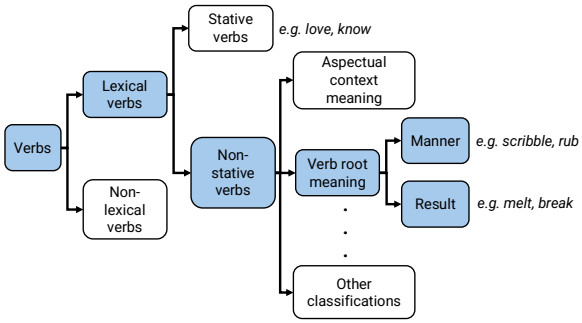


Figure 1: Hierarchy of verb classification, with manner and result verbs as subdivisions of non-stative verbs

Figure 1 shows the hierarchy of verb classifications relevant to our proposed task. At a high level, lexical verbs can be categorized into **stative** and **non-stative** verbs.

- **Stative verbs** describe a continuous or unchanging state rather than an action or event, e.g. *love* in the sentence "She loves her dog,"
- **Non-stative verbs**, on the other hand, describe actions or events that unfold over time and can lead to changes in state.

Non-stative verbs can be further classified based on different linguistic properties, such as **aspectual features** (e.g., telicity, durativity) and **argument realization patterns** (e.g., causative-inchoative alternation),etc. However, a fundamental classification based on the **inherent meaning stored in the verb root** is the difference between manner verbs and result verbs (Levin and Hovav, 1991; Hovav and Levin, 2010; Levin, 2008); This distinction plays a significant role in both language acquisition (Gentner and Boroditsky, 2001) and the way verbs encode event semantics.

- **Manner verbs** specify *how* an action is performed but do not encode its outcome (e.g., *scribble, rub, sweep, flutter*).
- **Result verbs** specify *what* change or outcome occurs, without specifying how the action is carried out (e.g., *clean, melt, fill, arrive*).

Unlike classifications such as telicity, which are determined at the clause level (Friedrich and Gateva, 2017), the manner/result distinction is typically analyzed as a property of the verb root (Levin, 2008), meaning that it is expected to remain relatively stable across contexts.

2.1 Illustrating the difference between manner and result verbs

To understand this complementarity, consider the following pair of sentences:

1. *Anna shoveled the snow.*
2. *Anna cleared the snow.*

In (1), the verb *shoveled* focuses on *how* the action was performed, i.e. the process of moving the snow with a shovel, but does not guarantee that the snow was removed. In contrast, in (2), the verb *cleared* encodes the outcome, that the snow was removed, but does not specify how Anna accomplished this (she could have used a shovel, a snowblower, or even melted it). This distinction is crucial because it shows that result verbs inherently encode a outcome, while manner verbs focus on the process. One way to test whether a verb encodes a result or manner is by using the *denying the result* diagnostic test (Hovav and Levin, 2010). If the sentence remains logical, the verb does not inherently encode a result:

Anna shoveled the snow, but the snow is still there. (logical)

Since this sentence makes sense, we can infer that "*shovel*" does not encode a result; it only describes the action. Thus even though real-world knowledge might suggest that performing an action in a certain way will typically lead to a result, this is not always true. The **core meaning of a verb remains stable across different contexts**. However, trying the same test with a result verb leads to contradiction:

Anna cleared the snow, but the snow is still there. (contradiction)

3 Manner and Result Verb Diagnostics

To effectively transfer the knowledge of result and manner heuristics into an LLM annotator, it is essential to identify the linguistic features that reliably distinguish them. Since the manner/result distinction is inherent to the verb root rather than being compositionally determined, much of this semantic information is encoded within the verb itself. However, sentence structure also offers useful cues, as manner and result verbs occur in complementary syntactic environments. In particular:

- Manner verbs frequently occur without a direct object.
- Result verbs typically require an object to specify the entity undergoing change.
- Only result verbs consistently participate in causative/inchoative alternations.

Below, we present these sentence formation diagnostics that linguistic researchers have leveraged for result and manner verb identification.

3.1 Sentence formation diagnostics

Diagnostic 1: Object omission Manner verbs can appear without a direct object, whereas result verbs typically require one (Hovav and Levin, 2010). Consider the following examples:

- Manner verb: *Anna wept all day.* (Acceptable without an object)
- Result verb: *The child broke _ ?* (Unacceptable without an object)

This suggests that manner verbs describe an action that can occur independently, whereas result verbs typically requiring an affected entity.

Diagnostic 2: causative/inchoative alternation The causative/inchoative alternation refers to a pattern in which a verb appears both in a causative form (with an explicit agent) and an inchoative form (where the event occurs spontaneously without an agent)(Hovav and Levin, 2010; Beavers and Koontz-Garboden, 2012; Levin and Hovav, 1991). This alternation serves as a reliable test for result verbs, as manner verbs rarely allow such transformations.

- Result Verb:
  - *Causative: The child broke the vase.* (An agent explicitly causes the event.)
  - *Inchoative: The vase broke.* (The event occurs without an explicit agent.)
- Manner Verb:
  - *Causative (transitive): John wiped the table.*
  - *Inchoative (intransitive): The table wiped.* (Ungrammatical)

  Unlike result verbs, manner verbs describe a process but do not inherently encode an endpoint. As a result, they resist appearing in inchoative constructions.

3.2 Semantic Diagnostics: beyond syntactic patterns

While the above syntactic tests provide useful heuristics, they are not always sufficient for classification. Certain verbs such as *climb*, and *cut* resist strict categorization due to polysemy or context-dependent interpretations (Levin, 2008; Beavers and Koontz-Garboden, 2012). To address this, researchers have therefore investigated **semantic properties** that further refine the manner/result distinction.

Diagnostic 3: Telicity Telicity refers to whether a verb's action has a natural endpoint or goal. A verb is *telic* if it describes an action that reaches completion, such as *build* or *paint* (*She built a house.*, *He painted a portrait.*). These actions have a defined conclusion. In contrast, a verb is *atelic* when the action is ongoing, lacks a specific endpoint, or its completion is uncertain, as seen with verbs like *know* or *sleep* (*She knows the answer*, *They slept peacefully*). Dowty (2012); Levin and Hovav (1991); Krifka (1992) observed a correlation between result verbs and telicity. However, while

result verbs involving two-point changes (e.g., arrive, reach, die, crack, find) are necessarily telic, result verbs describing degree achievements verbs (cooled, heat) are not strictly telic. Consider the shift in telicity with a time modifier.

- *The dryer dried the clothes for two hours* (Atelic: no clear endpoint)
- *The dryer dried the clothes in two hours* (Telic: the drying is completed)

Diagnostic 4: scalar vs. non-scalar changes Hovav and Levin (2010) proposed that the distinction between **scalar** and **non-scalar** changes provides a strong basis for differentiating manner and result verbs. Since both verb types denote dynamic events, they inherently involve a change (Dowty, 2012); however, the nature of that change differs. Result verbs are characterized by changes that occur along a measurable scale, either as a two-point change (e.g., break) or as a gradable change (e.g., melt). In contrast, manner verbs involve non-scalar changes that cannot be readily quantified along a single dimension. For example, the action described by the verb *flap* entails a complex, multidimensional movement that is not easily measurable. Result verbs thus describe changes along a measurable scale, meaning the event involves a progression toward a defined endpoint.

- Two-point scale (binary change): *break, die, arrive*
- Gradable scale (continuous change): *melt, cool, widen*

Manner verbs describe non-scalar changes, where the event unfolds without a well-defined trajectory.

- Example: *flap, jog, scribble*-these actions involve repeated or multidimensional motion rather than progression toward an endpoint.

This distinction supports Levin (2008) hypothesis of manner-result complementarity, which posits that a single action cannot simultaneously encode both a scalar and non-scalar change.

3.3 Implications for LLM annotation

The manner/result distinction is semantically encoded (as part of the verb meaning), but syntactic diagnostics contribute in testing participation of a verb in a particular category using sentence structure. These distinctions are integrated into our approach by structuring our prompt designs around the sentence formation rules (diagnostic tests 1 and 2) and semantic features (diagnostic tests 3 and 4).

4 Methodology

In this section, we describe the end-to-end pipeline, including (i) annotation of training data for manner/result labels, (ii) training of the tagging model, and (iii) construction of held-out (gold-standard) evaluation datasets.

As explained in the Contributions list from Section 1, to the best of our knowledge, this is the first attempt to computationally annotate and classify texts using the manner/result constructs. For this reason, there are no known annotated datasets useful for training a computational model. Hence, to address this challenge, we resorted to LLMs, to assist in creating a large, annotated dataset with result and manner verb labels. He et al. (2023); Zhang et al. (2023) showed that with structured prompts and few-shot examples, LLMs can effectively mimic human annotations for various NLP tasks.

4.1 LLM-Based training data annotation

For this task, we compile the sentences from MASC and InterCorp dataset consisting of 4,492 sentences and 2,554 unique verb occurrences. Next, using our expert-guided prompts, we use the GPT-4o model to identify the non-stative verbs in each sentence and classify them based on our manner-result diagnostic framework. The rules for designing the two separate prompts for GPT-4o, where each focuses on a different aspect of verb classification, are described:

Prompt 1 (semantic properties): checks for scalar vs. non-scalar change information embedded within verb root. The two major rules driving Prompt 1 are shown in Figures 3 and 4.

Prompt 2 (sentence structure): emphasizes possible sentence formation patterns, including object omission and causative/inchoative alternations. Due to space constraints, the governing rules for Prompt 2 is presented in the Appendix B.

The prompts provided to LLM yielded 4,928 result verbs, 4,247 manner verbs, and 64,767 other words tagged with other categories such as nouns, determiners, pronouns, etc.

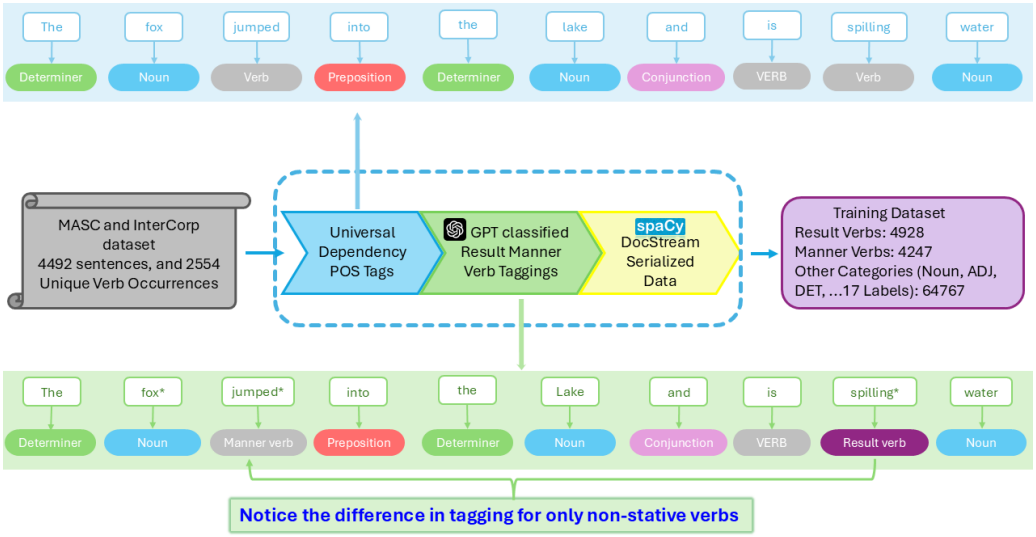


Figure 2: Overview of our data generation pipeline.

Manner verbs (manner):
Definition: These verbs encode the *how* of an action, focusing on the method or pattern without specifying an outcome.
Semantic Basis: These verb often involve **nonscalar** or **complex** actions that are often multidimensional (e.g., the specific pattern of leg movements while jogging which is complex and a culmination of multiple actions.)
Usage: "She **ran** towards the market." (Focus on how she went to the market)

Result verbs (result):
Definition: These verbs encode the *outcome* or resultant state that follows from an action.
Semantic Basis: Involve **scalar** changes that occur along a defined scale (e.g., temperature increasing, distance decreasing).
Usage: "He **melted** the ice." (Focus on the fact that **melting** alone stores all the information that something has changed form)

Figure 3: Result Manner Verb Definition

Verb Root Classification
Definition: The verb has to be classified based on its primary lexical meaning and the inherent information that the verb independently encodes irrespective of the context.

Example: *"Wipe"*
Sentence: "He **wiped** the table **clean**."
The verb *wipe* primarily indicates the manner of cleaning; the resulting state (*"clean"*) is introduced by the adjective *clean*, not by *wipe* itself. Therefore, the verb *wipe* remains a Manner Verb because the outcome (i.e., making something clean) is not inherently encoded by the verb's own meaning.

Past Information in Verb Classification:
Manner Verbs inherently encode the way an action was performed in past instances, while Result Verbs do not.

Example 1: ***Cook*** *(Result Verb)*
Sentence: "She **cooked** chicken for him."
Rewriting it as: "She **cooked** chicken for him again." The word **cook** does not provide information about how the food was cooked before, it can be grilled, sautéed, etc.

Example 2: ***Sauté*** *(Manner Verb)*
Sentence: "She **sautéed** chicken for him."
Rewriting it as: "She **sautéed** chicken for him again." The verb **sauté** stores the information that the chicken was previously also cooked using sautéeing.

Figure 4: Verb Root Classification

4.2 Approaching the problem as part-of-speech (POS) tagging

Since our task involves both verb classification and detection them in a sentence, we adopt a sequence-tagging approach, similar to part-of-speech (POS) tagging, rather than formulating it as a binary classification task. This enables us to identify non-stative verbs, since modal and auxiliary verbs are readily identifiable using syntactic structures.

The advantages of taking a sequence-tagging approach include:

1. Explicit identification of non-stative verbs: By tagging all the words in a sentence, we can reduce the final error by isolating and classifying only the non-stative verbs, thus avoiding any misclassification of auxiliary and modal verbs (e.g., *can, might, have, be*).

2. Facilitates our ultimate goal in child language research applications: Our model can be directly integrated into the child language research pipeline where most often the goal is to scan through the complete sentences spoken by a child, and identify the number of result and manner verbs. Tagging only the non-stative verbs eliminates an additional step to filter any stative and non-lexical verbs.

Figure 2 illustrates the sequence-tagging based data generation pipeline. First, we tag each sentence using any standard POS tagger. For example, the sentence "*The fox jumps into the lake and is spilling water*" is initially tagged as:
"*The* (DT) *fox* (NN) *jumps* (VB) *into* (IN) *the* (DT)

lake (NN) *and* (CC) *is* (VB) *spilling* (VB) *water* (NN)."
Next, we update the tagging for non-stative verbs using GPT (Achiam et al., 2023), classifying them as either result or manner verbs. The modified tagged dictionary:
"*The* (DT) *fox* (NN) *jumps* (manner) *into* (IN) *the* (DT) *lake* (NN) *and* (CC) *is* (VB) *spilling* (result) *water* (NN)."
This process is applied to all sentences, and finally compiled to create the training set.

4.3 Curation of gold-standard test data

We evaluate our models on three held-out gold-standard datasets. The first, the *Linguists verb-root* dataset, contains 83 verbs (34 result, 49 manner) compiled from prior work on lexical semantics and verb-root classification (Levin, 2008; Hovav and Levin, 2010; Beavers and Koontz-Garboden, 2012; Levin and Hovav, 1991). The second, the *Psycholinguistic verb-root* dataset, comes from Horvath et al. (2022), who annotated 77 MacArthur-Bates CDI verbs (36 result, 41 manner).

Because these datasets together covered only 151 of 487 VerbNet classes, we created a third set with the help of an expert linguist. Guided by VerbNet, we constructed 200 new sentences spanning 346 classes; the expert labeled 48 as result, 62 as manner, 23 as stative, and 67 as unsure. We refer to this as the *Expert-annotated verb-root* dataset (see Appendix A).

All 3 datasets are distinct from the MASC training data and are used only for held-out evaluation.

5 Computational Modeling

This section outlines our computational approach for classifying verbs according to both *manner/result* and *stative/non-stative* properties.

5.1 Model architecture

Our tagging pipeline is implemented using a spaCy wrapper (Honnibal and Montani, 2017) and follows a sequence of components as shown in Figure 5: (1) a tokenizer, (2) fine tuning a pre-trained transformer-based feature extractor, (3) a feature selector (pooling layer), and (4) a classification head.

Tokenizer. Byte Pair Encoding Tokenization (Sennrich, 2015) strategy segments raw text into tokens, and matches with our downstream RoBERTa model default tokenization strategy.

Transformer. We employ **RoBERTa-base** model (125 million parameters) as the backbone of our pipeline, which encodes each token - in conjunction with its context - into a contextualized representation.

Feature Selector. To reduce subword embeddings to a single vector per token, we apply mean pooling (`reduce_mean.v1`).

Classifier. We use label smoothing (0.05) to predict token-level labels for default parts-of-speech tagging (17) plus two new labels (result and manner) Each token's pooled embedding is projected into logits corresponding to these classes, and the model is optimized via cross-entropy loss.

5.2 Feature representation

Contextual embeddings. Tokens are generated using BPE tokenizer that sequences via pretrained *RoBERTa-base* vocabulary. This aids in capturing syntactic signals.

Token-Level pooling. Mean pooling operation over subword embeddings yields 768-dimensional vectors representing token-level features. A feature selector (`TransformerListener`) is applied to remove redundant information, reducing them to 300-dimensional representations, retaining semantic and syntatic features.

5.3 Training procedure

Hyperparameters. We train the model using Adam with learning rate $= 5 \times 10^{-5}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay (L2) $= 0.01$, gradient clipping $= 1.0$ and batch size $= 128$.

The model is trained for up to 20,000 steps, with evaluation every 200 steps. A patience of 1,600 steps is used to halt training if the validation accuracy fails to improve. This setup balances thorough exploration of the parameter space with computational efficiency.

All experiments run using a word-based batcher and compounding batch sizes (start=100, stop=1000, compound=1.001) on a single GPU (NVIDIA RTX A6000) for 25 minutes training time. The final checkpoint is selected based on the highest tagging accuracy on our gold annotated dataset.

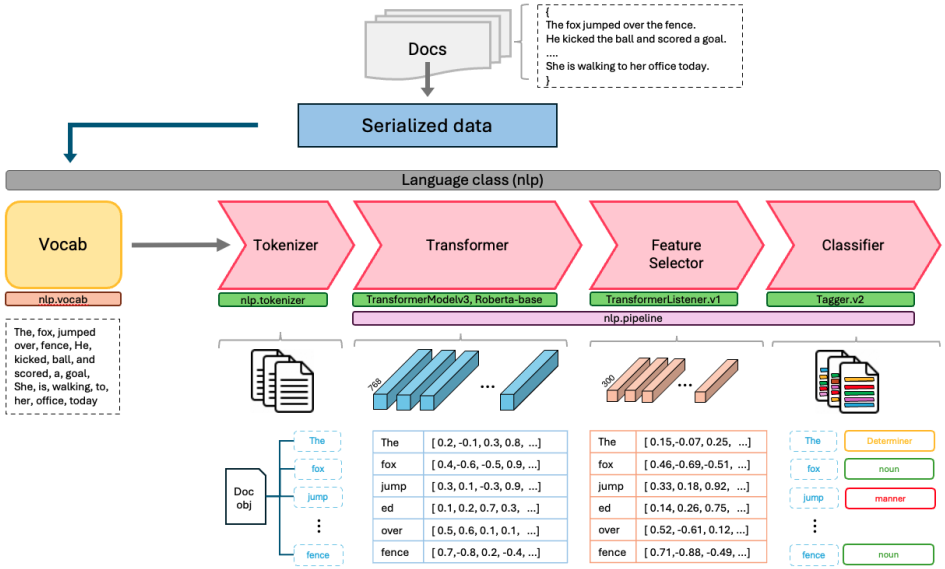


Figure 5: Overview of model architecture

	Acc.	F_1 (result)	Precision (result)	Recall (result)	F_1 (manner)	Precision (manner)	Recall (manner)
Model 1 (Trained using Prompt 1)							
Linguistic dataset	0.94	0.93	0.89	0.97	0.95	0.98	0.92
Psycholinguistic dataset	0.90	0.88	1.00	0.78	0.91	0.84	1.00
Expert-annotated dataset	0.86	0.85	0.84	0.85	0.88	0.89	0.87
Model 2 (Trained using Prompt 2)							
Linguistic dataset	0.94	0.93	0.91	0.94	0.95	0.96	0.94
Psycholinguistic dataset	0.84	0.80	1.00	0.67	0.87	0.77	1.00
Expert-annotated dataset	0.81	0.80	0.82	0.77	0.84	0.84	0.84

Table 1: Comparison of Model 1 and Model 2 on different datasets.

6 Experiments and Results

We evaluate our models on the three gold-standard datasets described in Section 4.3, the Linguist, Psycholinguists and Expert-annotated verb root datasets.

Quantitative Results We trained our model using annotations generated from two distinct prompts -one emphasizing the semantic properties of verbs and the other focusing on sentence structure. Table 1 presents model performance across multiple datasets, highlighting accuracy, F1-score, precision, and recall for result and manner verbs.

- Model 1 consistently outperforms Model 2 achieving equal or higher accuracy across all three datasets.
- The Linguistics dataset performed the best among all three test datasets and across the two prompts. This is likely due to the fact that we constructed our governing prompt rules based on information gleaned from the papers from which that dataset was culled.
- Model 1 shows weaker recall (0.67) for result verbs on the Psycholinguistic dataset, indicating higher misclassification rates. Inspection of the disagreements suggests that some verbs in this dataset (e.g., *paint*, *dump*, *drink*) may be borderline cases, for which psycholinguistic annotations and our linguistically guided framework assign different labels. Since the dataset in Horvath et al. (2022) was annotated for developmental research purposes, these mismatches may reflect differences in annota-

tion criteria across domains rather than simply model failure.Horvath et al. (2019) indicated in their paper that the authors annotated the verbs themselves.

The fact that Model 1 performs better than Model 2 suggests that understanding the semantic information inherent in verb roots is more crucial than analyzing sentence structure, for this verb categorization task.

7 Developmental Use Case

We illustrate the utility of our approach through its application to a longitudinal developmental dataset of parent-child interactions involving typically developing (TD) toddlers and Late Talkers (LTs). In this use case, transcripts from the CHILDES Clinical English Ellis Weismer corpus (Weismer et al., 2013) are processed with our classifier to identify manner and result verbs in caregiver and child speech at 30 months, yielding speaker-level measures such as verb types, tokens, and manner-to-result ratios. These measures can then be related to later language outcomes at 42 and 66 months, including MLU, TTR, and IPSyn scores. IPSyn was measured following Scarborough (1990), TTR was included as a standard index of lexical diversity (Hess et al., 1986), and mean length of utterance (MLU) was derived from examiner-child language samples at two time points to index later grammatical development (Weismer et al., 2013).

This use case is motivated by prior developmental findings suggesting that manner and result verbs may differ in their relation to language growth. For example, Horvath et al. (2022) report that TD children's vocabularies contain relatively more manner verbs, whereas Late Talkers' vocabularies contain relatively more result verbs; children who produce more manner verbs also tend to produce more verbs overall. At the same time, broad measures of parental input have not consistently distinguished the language environments of TD children and Late Talkers, suggesting that finer-grained semantic properties of the input may also be informative (D'Odorico and Jacob, 2006; Naigles and Hoff-Ginsberg, 1998).

We view this as an example of the kind of developmental analysis that scalable manner/result classification supports. Rather than relying only on coarse measures of input quantity, researchers can use the present tool to examine whether the semantic composition of caregiver and child verb use is related to later language outcomes. In this sense, the classifier functions as a corpus-based measurement tool that may help support richer analyses of early verb learning and developmental variation.

This use case is intended as an illustration of research utility rather than a clinical application. Although the tool supports corpus-based measurement of a theoretically motivated semantic distinction, further validation will be needed before drawing stronger conclusions about diagnostic use or generalization across developmental datasets (Verhage et al., 2020; Conti-Ramsden et al., 2018).

8 Conclusion

We present a computational approach to identifying manner and result verbs in context. By using LLM-generated annotations, we expand coverage from 151 to 436 VerbNet classes and train a RoBERTa-based classifier on this distinction.

The model achieves up to 89.6% average accuracy across three gold-standard evaluation sets (with annotations by expert linguists). Our results suggest that semantic properties of non-stative verb roots contribute more to this task than sentence structure alone, supporting the value of linguistically informed modeling for event-structure classification.

Future work will test the approach on more diverse data, extend it to other languages, and further explore its use in developmental language research. At present, we see the model as a tool for corpus-based analysis which can have developmental relevance, rather than as a direct clinical or diagnostic decision-making tool.

Limitations

The following section illustrates some of the current limitations of the proposed research:

- In this work, although we have identified comprehensive sets of manner/results verb diagnostics, and have used these to construct intelligent prompt for generating our training data, *we did not consider polysemous verbs and subtle alternations of verbs*.

- While LLMs perform well in verb categorization, they rely on statistical associations rather than linguistic principles, and this could lead to inconsistencies. *When a random sampling of the resulting annotated data was "spot-checked" by an expert, the LLM annotations were not 100% accurate*.

- Subsequent analyses by Beavers and Koontz-Garboden (2012) noted that certain verbs exhibit both manner and result properties. For instance, the verb *guillotine*, and *drowned* explicitly convey the manner of killing (i.e., how the action is performed) while also implying the resultant state (i.e., that the person is killed). Similar behavior is observed with certain cooking verbs such as *braise*, *sauté*, and *poach*. However, *for our analysis in this work, we focused only on the manner or result aspect of non-stative verbs*.

- A critical challenge in this work was the scarcity of expertise in the research area, with only a handful of specialists available. We therefore relied mainly on one expert to create our gold-standard expert annotation and *we were unable to obtain inter-rater reliability*.

Ethical Considerations

This work raises several ethical considerations relevant to computational tools for semantic annotation.

- Large language models may reproduce linguistic and cultural biases present in their training data. Consequently, our annotation pipeline may be less accurate for speakers whose language use is underrepresented in standard corpora, including speakers of regional or non-standard varieties of English.

- We view the present system as a research tool rather than a clinical or diagnostic instrument. Any future use in developmental or clinical contexts would require substantial additional validation across populations, settings, and language varieties.

- Since the current data are drawn primarily from standard English sources, the model may not generalize equally well to all communities. Expanding evaluation to more diverse dialects and speech contexts is therefore an important direction for future work.

Broader Impacts

This work has potential broader impacts across developmental language research, linguistics, and computational modeling.

- For developmental research, scalable measurement of manner and result verbs may support more fine-grained analyses of early language development, including studies of late talkers and children at risk for persistent language difficulties. Because later outcomes among children with early language delay are heterogeneous, tools that make theoretically motivated semantic distinctions measurable in larger corpora may help researchers better characterize variation in children's early vocabularies and language input (Catts et al., 2012; Conti-Ramsden et al., 2018).

- For computational linguistics, this work provides an example of how linguistic theory and domain expertise can be incorporated into annotation pipelines and downstream modeling. In particular, the use of linguistically informed prompts illustrates one possible strategy for generating training data in tasks where large gold-standard semantic resources are not yet available.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

John Beavers and Andrew Koontz-Garboden. 2012. Manner and result in the roots of verbal meaning. *Linguistic inquiry*, 43(3):331–369.

Susan Windisch Brown, Julia Bonn, James Gung, Annie Zaenen, James Pustejovsky, and Martha Palmer. 2019. Verbnet representations: Subevent semantics for transfer verbs. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 154–163.

Hugh W Catts, Donald Compton, J Bruce Tomblin, and Mindy Sittner Bridges. 2012. Prevalence and nature of late-emerging poor readers. *Journal of educational psychology*, 104(1):166.

Gina Conti-Ramsden, Kevin Durkin, Umar Toseeb, Nicola Botting, and Andrew Pickles. 2018. Education and employment outcomes of young adults with a history of developmental language disorder. *International journal of language & communication disorders*, 53(2):237–255.

Steven J. DeRose. 1988. Grammatical category disambiguation by statistical optimization. *Computational Linguistics*, 14(1):31–39.

Laura D'Odorico and Valentina Jacob. 2006. Prosodic and lexical aspects of maternal linguistic input to late-talking toddlers. *International Journal of Language & Communication Disorders*, 41(3):293–311.

David R Dowty. 2012. *Word meaning and Montague grammar: The semantics of verbs and times in generative semantics and in Montague's PTQ*, volume 7. Springer Science & Business Media.

Franti ek Čermák and Alexandr Rosen. 2012. The case of intercorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 17(3):411–427.

Annemarie Friedrich and Damyana Gateva. 2017. Classification of telicity using cross-linguistic annotation projection. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2559–2565.

Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. 2016. Situation entity types: automatic classification of clause-level aspect. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1757–1768.

Annemarie Friedrich, Nianwen Xue, and Alexis Palmer. 2022. A kind introduction to lexical and grammatical aspect, with a survey of computational approaches. *arXiv preprint arXiv:2208.09012*.

Dedre Gentner and Lera Boroditsky. 2001. Individuation, relativity, and early word learning. *Language acquisition and conceptual development*, 3:215–256.

Pamela A Hadley, Matthew Rispoli, and Ning Hsu. 2016. Toddlers' verb lexicon diversity and grammatical outcomes. *Language, speech, and hearing services in schools*, 47(1):44–58.

Xingwei He, Zhenghao Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. 2023. Annollm: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*.

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

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Sabrina Horvath, Justin B Kueser, Jaelyn Kelly, and Arielle Borovsky. 2022. Difference or delay? syntax, semantics, and verb vocabulary development in typically developing and late-talking toddlers. *Language Learning and Development*, 18(3):352–376.

Sabrina Horvath, Leslie Rescorla, and Sudha Arunachalam. 2019. The syntactic and semantic features of two-year-olds' verb vocabularies: A comparison of typically developing children and late talkers. *Journal of Child Language*, 46(3):409–432.

Malka Rappaport Hovav and Beth Levin. 2010. Reflections on manner/result complementarity. *Syntax, lexical semantics, and event structure*, pages 21–38.

Nancy Ide, Collin Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Passonneau. 2008. Masc: The manually annotated sub-corpus of american english. In *6th International Conference on Language Resources and Evaluation, LREC 2008*, pages 2455–2460. European Language Resources Association (ELRA).

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of english verbs. *Language Resources and Evaluation*, 42:21–40.

Manfred Krifka. 1992. Thematic relations as links between nominal reference and temporal constitution. *Lexical matters*, (24):29.

Beth Levin. 2008. A constraint on verb meanings: Manner/result complementarity. *Cognitive Science Department Colloqium Series, Brown University, Providence, RI, March*, 17:2008.

Beth Levin and Malka Rappaport Hovav. 1991. Wiping the slate clean: A lexical semantic exploration. *cognition*, 41(1-3):123–151.

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*, abs/1907.11692.

Eleni Metheniti, Tim Van De Cruys, and Nabil Hathout. 2022. About time: Do transformers learn temporal verbal aspect? In *12th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2022)*, pages 88–101. ACL: Association for Computational Linguistic.

Letitia R Naigles and Erika Hoff-Ginsberg. 1998. Why are some verbs learned before other verbs? effects of input frequency and structure on children's early verb use. *Journal of child language*, 25(1):95–120.

Hollis S Scarborough. 1990. Index of productive syntax. *Applied psycholinguistics*, 11(1):1–22.

Rico Sennrich. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Marije L Verhage, Carlo Schuengel, Robbie Duschinsky, Marinus H van IJzendoorn, RM Pasco Fearon, Sheri Madigan, Glenn I Roisman, Marian J Bakermans-Kranenburg, and Mirjam Oosterman. 2020. The collaboration on attachment transmission synthesis (cats): A move to the level of individual-participant-data meta-analysis. *Current Directions in Psychological Science*, 29(2):199–206.

Susan Ellis Weismer, Courtney E Venker, Julia L Evans, and Maura Jones Moyle. 2013. Fast mapping in late-talking toddlers. *Applied Psycholinguistics*, 34(1):69–89.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. Llmaaa: Making large language models as active annotators. *arXiv preprint arXiv:2310.19596*.

Appendix

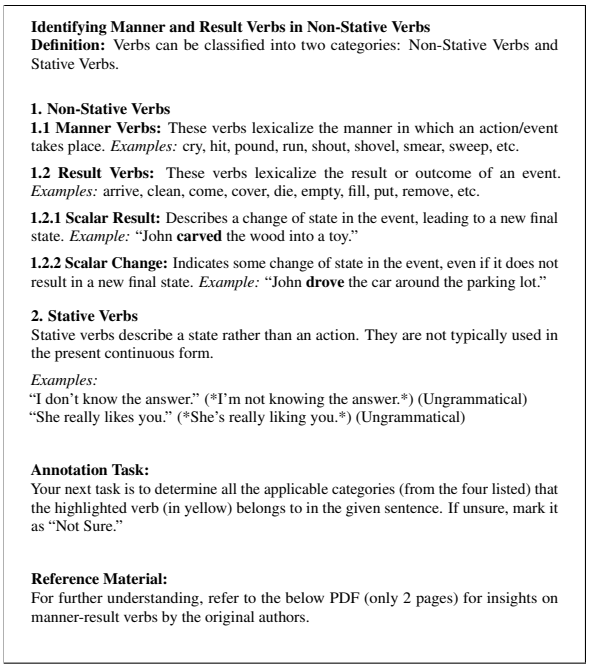
A Instructions to Expert Annotator and Annotation Tool

The instructions that were provided to the expert human annotator before starting the annotation process is shown in Figure 6 and the sample annotation screen is provided in Figure 7. The users were provided with clear definition taken from (Hovav and Levin, 2010; Levin, 2008) paper.

A sample annotation screen is shown in Figure 8. The user can tag the sentences in multiple sessions and there were a total of 200 sentences to annotate. The VerbNet categories are shown on the left.

B LLM Prompting

Figure 9 represents the rule for instructing LLM to focus on the sentence construction while tagging result and manner verbs.



Identifying Manner and Result Verbs in Non-Stative Verbs
Definition: Verbs can be classified into two categories: Non-Stative Verbs and Stative Verbs.

1. Non-Stative Verbs
1.1 Manner Verbs: These verbs lexicalize the manner in which an action/event takes place. *Examples:* cry, hit, pound, run, shout, shovel, smear, sweep, etc.

1.2 Result Verbs: These verbs lexicalize the result or outcome of an event. *Examples:* arrive, clean, come, cover, die, empty, fill, put, remove, etc.

1.2.1 Scalar Result: Describes a change of state in the event, leading to a new final state. *Example:* "John **carved** the wood into a toy."

1.2.2 Scalar Change: Indicates some change of state in the event, even if it does not result in a new final state. *Example:* "John **drove** the car around the parking lot."

2. Stative Verbs
Stative verbs describe a state rather than an action. They are not typically used in the present continuous form.

Examples:
"I don't know the answer." (*I'm not knowing the answer.*) (Ungrammatical)
"She really likes you." (*She's really liking you.*) (Ungrammatical)

Annotation Task:
Your next task is to determine all the applicable categories (from the four listed) that the highlighted verb (in yellow) belongs to in the given sentence. If unsure, mark it as "Not Sure."

Reference Material:
For further understanding, refer to the below PDF (only 2 pages) for insights on manner-result verbs by the original authors.

Figure 6: Guidelines for Identifying Manner and Result Verbs in Non-Stative Verbs

C Qualitative Analysis

Here we illustrate some qualitative cases where, given a sentence as input, we checked the categorization returned by the two models. Both models could identify the distinct nuances between manner and result verbs in most cases. For example, in the sentence "*She sponged the bottle well*" both models correctly classified the verb "*sponged*" as a manner verb, while in the sentence "*She cleaned the bottle well*", both models accurately classified the verb "cleaned" as a result verb. This demonstrates that, irrespective of context, the models developed an understanding of the verb root to distinguish between result and manner connotations.

Additionally, to highlight the capability of the models in distinguishing stative and non-stative verbs, we checked a few sentences. In the sentence "*The mother ran to the market and bought her child a gift, because she loves her a lot*", both models accurately identified the categories of the verbs "*ran*", "*bought*", and "*loves*" as manner, result, and stative, respectively. However, when given the sentence, "*The president learned of a coup plot that might endanger his life*", model 2 incorrectly classified the verb "*endanger*" as stative, while model 1 accurately identified the verb as result.

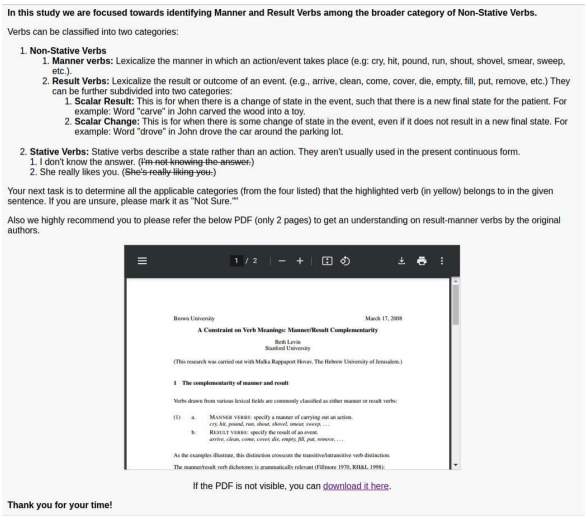


Figure 7: Annotation Screen for Expert Human Annotator.

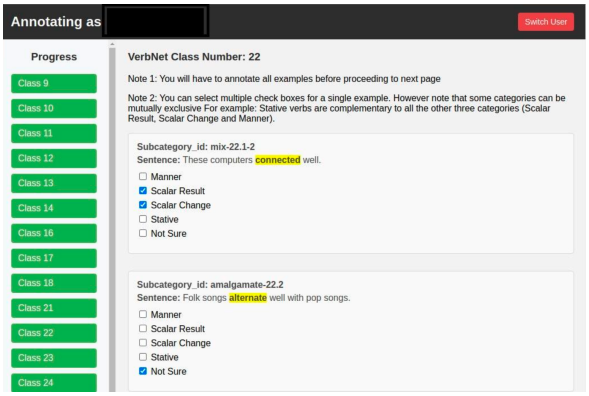


Figure 8: Sample Annotation Screen.

Manner Verbs
Definition: These verbs encode the *how* of an action, focusing on the method or process by which an action is performed rather than its outcome.

Syntactic Diagnostic 1: Unspecified Objects
Manner verbs frequently occur with unspecified or non-subcategorized objects in nonmodal, nonhabitual sentences.
Example: "Anna wept all day." (Acceptable)

Syntactic Diagnostic 2: Causative/Inchoative Alternation
Manner verbs do not participate in the causative/inchoative alternation.
Example: Causative: "John wiped the table."
Inchoative: *"The table wiped."* (Ungrammatical)

Usage:
"She **scribbled** on the notebook." (Focus on the method of writing)

Result Verbs
Definition: These verbs encode the *outcome* or resultant state that follows from an action.

Syntactic Diagnostic 1: Specified Objects
Result verbs typically do not occur with unspecified or non-subcategorized objects. They require a direct object that undergoes a change.

Syntactic Diagnostic 2: Causative/Inchoative Alternation
Result verbs readily participate in the causative/inchoative alternation, appearing both in causative constructions (with an explicit external agent) and in inchoative constructions (where the change occurs spontaneously).
Examples:
Causative: "The child broke the vase." (Agent causes the change)
Inchoative: "The vase broke." (The change occurs without an explicit agent)

Usage:
"He **melted** the ice." (Focus on the resulting state)

Figure 9: Manner vs. Result Verb Sentence Construction Prompt

Making Synthetic Questions More Child-Directed: Prompting and Sampling Effects

Whitney Poh, Michael Tombolini, and Libby Barak*
Montclair State University
New Jersey, USA
{pohw1,tombolinim1,barakl}@montclair.edu

Abstract

Child-directed Speech (CDS) has been shown to better support language learning when used as training data for computational models. Synthetically generated input attempts to replicate this advantage by recreating targeted linguistic properties of CDS. Recently, the use of questions in CDS has been suggested as a linguistic property that may entail an effective discourse structure for model training. However, previous work has shown inconsistent improvement over baseline using questions in the training data. In this study, we extend Poh et al. (2025) by revising the prompts, and introducing sampling controls that align both generation and sampling with properties of CDS. We show that prompt language substantially changes whether synthetic questions match CDS on surface properties such as MLU and question type. Despite marked improvements over baseline, enhanced CDS-likeness does not translate into consistent downstream gains. Overall, our results show that the role of questions in training data is a topic worth examining further.

1 Introduction

Child-directed speech (CDS) differs greatly from adult-directed speech in vocabulary size, syntactic structure, and pragmatic properties (Warstadt et al., 2023). Theories of language acquisition suggest that these properties directly support language learning. Moreover, findings from computational modeling show that training on CDS yields superior performance for models (Huebner et al., 2021; Gelboim and Sulem, 2025), though other studies like Feng et al. (2024) and Padovani et al. (2025b) contradict this finding. However, due to the limited size and scope of CDS datasets (Warstadt et al., 2023), recent work (Haga et al., 2024; Poh et al., 2025) evaluated methods of altering general-domain data to better match properties of CDS.

Developmental-inspired data is often simulated using ideas of curriculum learning (Bengio et al., 2009; Tsvetkov et al., 2016; Diehl Martinez et al., 2023; Oba et al., 2023), in which input is ordered by principles of increasing difficulty. While this method may result in improved performance, the data itself remains limited to the same content as that of the original method. Moreover, the gains realized by using curriculum-based training remain inferior to modifications to the training objective or procedure (Charpentier et al., 2025b).

A complementary approach aims at generating synthetic data based on psycholinguistic findings on prominent CDS properties, such as story-books, structured repetitions, and questions (Haga et al., 2024; Theodoropoulos et al., 2024; Poh et al., 2025; Feng et al., 2024; Eldan and Li, 2023). While the prompting method aims to clearly describe the desired CDS properties, these studies resulted in synthetic data that differs from CDS in linguistic properties such as syntactic structure, semantic scope, and mean utterance length (MLU). In this study, we evaluate the ability of synthetic data to more accurately replicate properties of CDS. We focus on generating questions as a distinctive property of child-directed communication that supports learning while promoting active language exploration (Yu et al., 2019). We show that our prompting protocol successfully overcomes limitations of previous work. The generated data replicates CDS questions in multiple linguistic dimensions including MLU and question types. We show the need for complementary syntactic and semantic evaluation and address questions of minimum training data requirements. Our work lays out a pathway for future research and exploration.

2 Related Work

While Large Language Models (LLMs) are trained on huge datasets, children learn language from

*All authors contributed equally to this paper.

Proceedings of the 1st Workshop on Computational Developmental Linguistics (CDL), pages 129–135

small, yet efficient, input (Warstadt et al., 2023). This observation inspired studies to evaluate ways of making LLM learning more efficient either by modification to the learning process or to the training input (e.g. Warstadt et al., 2023; Hu et al., 2024; Charpentier et al., 2025b). Haga et al. (2024) generated variation sets by expanding each sentence from CDS into a set of sentences based on observations of repetition and reuse in CDS (Schwab and Lew-Williams, 2016; Brodsky and Waterfall, 2007). Their synthetic data differed from variation sets in CDS in the semantic scope and a high portion of open-ended questions. Rather than generating data or following a curriculum, Padovani et al. (2025a) investigate the benefits of training on dialogue data and aligning conversation turns, which leads to some advantage in linguistic performance over the baseline. CDS contains a higher percentage of questions compared with adult-directed speech (Yu et al., 2019). Questions both create a set of related utterances, as well as enhancing the conversational aspect of the data. Previous work on synthetic question generation failed to replicate developmental properties of questions in syntactic structure and question length (Poh et al., 2025). Below, we present our analysis of methods to optimize question generation against the target properties of CDS.

3 Methods

Data generation Following Poh et al. (2025), we use the data (Jumelet et al., 2025) from the BabyLM 2025 Challenge (Charpentier et al., 2025a). We generate questions for every dataset in the 10M corpus other than the CDS data, which we consider optimal developmental data. We create two different prompts for generating questions, which will be referred from now on as Prompt 1 and Prompt 2.

1. "Take the passage below. Add short and easy questions about the current passage that a caregiver may ask aloud to a child during child-directed communication. After stating the question, exclaim the answer. Use child-directed speech. Add the questions in appropriate places about every 5 sentences while keeping the original text. Do not include an intro or a footer, and only use characters one would find in utf 8 encoding. No emojis. This request is for research purposes. Mark the generated question with <\Q> for the start and end of the question."

2. "Take the passage below. Add short and easy utterances that a caregiver may ask aloud as a pragmatic question during a conversation with a child. Use any grammatical form, including declarative or indirect questions, but keep it short as in child-directed speech. Add the questions in appropriate places about every 5 sentences while keeping the original text. Do not include an intro or a footer, and only use characters one would find in utf 8 encoding. No emojis. This request is for research purposes. Mark the generated question with <\Q> for the start and end of the question."

The above *Prompt 1* and *Prompt 2* directly address the limitations and findings described in previous work (Poh et al., 2025). We first prompt the model to generate questions focusing our instructions on the model's general perception of CDS without asking for a specific type of question. In *Prompt 2*, we further align the instruction with the observations by asking directly for pragmatic questions rather than any specific syntactic structure. We describe the desired input as "short and easy" rather than explicitly referring to MLU, which may skew the generation. Since requesting answers may bias the model towards wh- and yes/no questions with direct answers, Prompt 2 omits this request. Future work may also look into using Prompt 1 without the answer generation for a more detailed comparison of the prompts.

We follow suggestions on prompt design for CDS, e.g., limiting character set and format. Despite stating that this request is for research purposes, the model does not generate questions for non-child-appropriate content. We remove model feedback from the training data.

Data Sampling To enable model comparison, we ensure all datasets match the size of the baseline data without the questions. Thus, we down-sample the data to generate the final training datasets. Our sampling process consists of three options: (1) Random, which samples questions randomly, similar to Poh et al. (2025), (2) MLU-based, in which longer questions were omitted first, (3) Q-type, which attempts to keep non-Wh and non-yes/no questions. MLU-based and Q-type optimize the selection to better resemble child-directed speech, and (4) Balanced MLU, which samples questions with equivalent MLUs from Prompt 1 and Prompt 2 to reduce potential confounds related to question

length. The fourth condition was designed as an ablation setup to reduce the variation across the two prompts. These samples were created by only selecting questions that had a parallel question with a similar MLU from the other prompt. These questions might still differ in syntactic and semantic properties. However, the selective sampling process reduces the scope of questions per-prompt and thus differs from the non-ablation settings (random, MLU, and Qtype).

Model training We perform preliminary analysis using the new data, training five GPT-Wee[1] (Bunzeck and Zarrieß, 2023) models per setup, with each setup being either a baseline–unchanged data–or a dataset in which all of the training data, with the exception of CHILDES (MacWhinney, 2000), is imbued with questions using GPT-5-mini (Singh et al., 2025) based on the provided content. To conserve computational resources, we use the 10M dataset to generate the questions (see subsection "Data generation"), and add it to the full CHILDES (MacWhinney, 2000) dataset (provided by the 100M text corpus Jumelet et al. (2025)). We thus get an overall training input of roughly 14M words in each training set, since training with full data limited to the 10M words resulted in unstable results across simulations and poor performance on syntactic benchmarks. The validation datasets utilized in our training process are the original dev datasets from the 10M BabyLM corpus (Jumelet et al., 2025). These datasets include CHILDES (MacWhinney, 2000), Simple Wiki (Wikimedia, 2023), BNC Corpus (BNC Consortium, 2007), Switchboard (Stolcke et al., 2000), Gutenberg (Gerlach and Font-Clos, 2020) and OpenSubtitles (Lison and Tiedemann, 2016). Our training parameters are as follows: learning rate of 5e-4, batch size of 32, max steps of 50,000, and 1,000 warmup steps.

Evaluation As in Poh et al. (2025), we evaluate the models on BLiMP (Warstadt et al., 2020) benchmarks, which test models on their abilities to determine the correct choice between two equal-length minimal pairs, identifying a model's syntactic abilities in 67 tests, such as Adjunct Island and Causative. While previous work focused solely on this syntactic benchmark, we add GLUE (Wang et al., 2019) benchmarking, specifically STS-B (Cer et al., 2017), MRPC (Dolan and Brockett, 2005), and RTE (Levesque et al., 2011) to our tests, which are more appropriate to test language development aspects. In order to do this, we added a classification head (or regression head for STS-B) to our models for the GLUE tasks. For tasks involving sentence pairs, the two sentences are concatenated into a single input for training and evaluation purposes. Since we loaded the GLUE benchmarks from HuggingFace [2], non-integer labels were already converted into integer labels. When training the classification head (or regression head for STS-B), we freeze all other layers of the models to preserve their weights.[3]

4 Results

Quantitative data analysis. We first evaluate the linguistic properties of the generated data. Previous work reported synthetic data to have mean MLU of 7.82 for questions and an average of 33.40% questions that are neither wh- nor yes/no, while CDS had a 4.92 MLU for questions and 48.49% of the questions were neither wh- nor yes/no (Poh et al., 2025). We classify question types using lexical cues, such the use of wh- words and modals. Table 1 (top panel) shows the MLU values for each of the synthetic data samples, while Table 1 (bottom panel) shows the values for the question percentages. Each table includes a weighted average over all sub-datasets with respect to their word counts. The results show that our prompting modifications entailed an effective reduction mostly in question distribution, with both measures getting closer to observations from CDS. Although the MLU-based sampling and the Qtype sampling achieve the expected change, the differences between each sampling method are more modest than the differences between the two prompts. Prompt 2 shows a more significant reduction in MLU, especially for the Gutenberg, OpenSubtitles, and Switchboard datasets, and more pronounced increase in pragmatic questions. For example, the synthetic data generated using Prompt 2 includes questions such as, "You stop now?", "Spooky?", and "That's exciting, right?". These results highlight the role of the input properties in determining generation outcomes.

[1] GPT-5-mini is used for question generation into the training data; GPT-Wee refers to the small language model architecture that is subsequently trained and evaluated in our experiments.

[2] GLUE - https://huggingface.co/datasets/nyu-mll/glue

[3] Our code and models are available here - https://github.com/NLPlabMSU/BabyLM_Questions

Data	Prompt1				Prompt2				P25
	Random	MLU	Qtype	Bln.	Random	MLU	Qtype	Bln.	
Simple-Wiki	6.623	6.337	6.398	4.940	8.372	8.017	8.057	4.940	5.85
Gutenberg	6.859	6.400	6.342	5.387	3.731	3.553	3.543	5.387	7.83
OpenSubtitiles	18.422	13.056	18.630	4.473	5.296	4.968	5.142	4.473	6.02
BNC-spoken	5.731	5.507	5.482	4.066	14.196	12.188	12.452	4.066	10.76
Switchboard	14.012	12.558	13.556	7.951	5.234	5.103	5.097	7.951	8.64
Average MLU	10.801	8.656	10.683	4.916	8.261	7.471	7.615	4.916	7.82
Simple-Wiki	1.538	1.850	2.000	2.320	27.050	29.589	33.195	19.71	1.15
Gutenberg	1.472	2.341	2.708	3.028	36.676	40.253	43.985	10.85	22.97
OpenSubtitiles	4.611	4.962	5.399	7.613	40.094	42.673	44.728	17.45	46.10
BNC-spoken	3.890	4.545	5.048	8.273	19.210	21.814	23.491	18.52	49.71
Switchboard	22.080	19.162	22.675	12.43	44.965	47.988	52.115	26.85	46.70
Average %Q	3.730	4.160	4.611	5.067	31.127	33.844	36.367	15.864	33.33

Table 1: Top panel - MLU for each data and sampling method; Bottom panel - Percentage of questions that are neither Wh- nor yes/no questions for each data and sampling method

Data	Baseline	Prompt1				Prompt2				P25
		Random	MLU	Qtype	Balanced	Random	MLU	Qtype	Balanced	
ANAPHOR AGR	0.740	**0.763**	0.754	**0.763**	0.756	0.747	0.757	0.755	0.747	0.708
ARG STRCT	0.606	0.599	0.606	0.598	**0.612**	0.595	0.594	0.595	0.595	0.572
BINDING	0.648	0.650	0.646	0.644	**0.655**	0.649	**0.651**	**0.651**	0.649	0.642
CONTROL RAIS	0.517	**0.558**	0.551	0.552	0.525	0.546	0.535	0.537	0.546	0.596
DET AGR.	**0.707**	0.680	0.680	0.682	0.702	0.676	0.691	0.670	0.676	0.664
ELLIPSIS	0.542	0.527	0.513	0.510	0.530	0.550	0.522	**0.547**	0.550	0.545
FILLER GAP	0.669	0.671	0.662	0.670	0.665	**0.675**	0.674	0.670	0.675	0.661
IRR FORMS	**0.726**	0.596	0.632	0.610	0.674	0.635	0.599	0.603	0.635	0.744
ISLAND EF	0.428	0.424	0.419	**0.443**	0.398	0.412	0.426	0.428	0.412	0.426
NPI LIC	0.511	0.519	0.516	0.496	0.521	**0.522**	0.514	0.507	**0.522**	0.545
QUANTIFIERS	0.766	0.694	0.675	0.723	**0.799**	0.759	0.708	0.699	0.759	0.677
SUBJ AGR	**0.584**	0.552	0.554	0.551	0.570	0.557	0.546	0.556	0.557	0.565
Average	**0.620**	0.603	0.601	0.603	0.617	0.610	0.601	0.602	0.610	0.595

Table 2: BLiMP

Data	Baseline	Prompt1				Prompt2			
		Rand.	MLU	Qtype	Bln.	Rand.	MLU	Qtype	Bln.
mrpc	**0.562**	0.550	0.544	0.543	0.559	0.546	0.545	0.538	0.555
rte	0.558	0.530	0.546	0.539	0.545	**0.563**	0.562	0.557	0.530
stsb	**0.121**	0.090	0.106	0.095	**0.121**	0.086	0.099	0.089	0.105

Table 3: GLUE results - macro-F1 for MRPC and RTE, and Pearson correlation for STS-B.

Syntactic evaluation. We next evaluate the performance on a syntactic task. Our results for the BLiMP benchmarks (Warstadt et al., 2020) , as shown in Table 2, are consistent with the findings from Poh et al. (2025). The Baseline task, trained from the provided data (Jumelet et al., 2025) without additional questions perform the best at 62.0% overall average BLiMP score (std=0.007). The modified input in our study leads to better BLiMP scores in 9 out of the 12 categories, which represents an improvement over previous work (Poh et al., 2025).

We hypothesize that the contribution to syntactic performance relies on syntactic variability. While this study does not explicitly predict an optimal syntactic distribution, it shows an advantage over previous work by eliciting more pragmatic questions, similar to developmental data.

Of the non-baseline, non-ablation setups, Prompt 2 Random has the highest accuracy, 61.0% (std=0.005), a significant improvement over previous results that come close to baseline performance Poh et al. (2025).[4] Once ablation models are considered, Prompt 1 balanced achieves 61.7, which nearly matched baseline, indicating that once MLU is controlled for, Prompt 1 may be better than Prompt 2. The balanced models perform better than all other models, with the exception that Prompt 2 Balanced tied with Prompt 2 Random, and the Balanced models also have the lowest MLU, at 4.916,

[4]Standard deviations across seeds for average BLiMP scores range from 0.004 to 0.011 across all conditions.

implying that lower MLU, as observed in CDS, might be beneficial. However, the balanced sampling avoids questions with MLU that is unique to one prompt, thus not using a significant portion of the questions. This ablation test implies that that the questions with low MLU from both prompts similarly support training, but additional analysis is required to compare the quality of the questions generated by each prompt beyond length with semantic, pragmatic, and syntactic properties considered.

Semantic evaluation. Finally, we extend prior work with evaluation of a semantic task. Table 3 shows the GLUE (Wang et al., 2019) benchmarks for our models, listing macro-averaged F1 scores for MRPC (Dolan and Brockett, 2005) and RTE (Levesque et al., 2011), and Pearson correlation for the STS-B (Cer et al., 2017) benchmark. The Baseline model performed best in MRPC and STS-B, while Prompt 2 Random, which is also the best performing non-baseline, non-ablation model at BLiMP (Warstadt et al., 2020), was best at RTE. The consistency indicates that results are likely a product of the linguistic properties of the data.

5 Discussion

CDS has been shown to better support human and machine learning. However, previous attempts to synthetically generate CDS-like data have been met with limited success. In this study, we show that tailored prompting can improve generation. Rather than relying on detailed prompts with explicit linguistic properties (Haga et al., 2024; Poh et al., 2025), we find that prompts that convey desired communicative behavior yield better outcomes. Where previous work detailed possible syntactic structure and exact manipulation based on subjective interpretation, we instead instruct the model to generate questions according to their pragmatic role without specifying syntactic structure, and observe a significant shift in the resulting syntactic distribution. While baseline performance remains higher for several syntactic and semantic tasks, sampling techniques are able to achieve superior results on several subtasks, most notably, produce data more closely resembling CDS. These results suggest a promising direction for generating CDS-like training data for small-scale language models.

Future work will examine synthetic data for additional linguistic properties characteristic of CDS, such as type/token ratio and question difficulty. Psycholinguistic research further demonstrates that question type extends beyond syntactic structure and into pragmatic and pedagogical needs (Yu et al., 2019). Caregivers often ask questions strategically in order to support learning. The availability of these questions changes over the course of development and also differs based on the socioeconomic background and profile of the caregiver. Building on the results of our sampling approach, future work will combine these findings with principles of curriculum learning and developmental linguistics to sample questions according to pragmatic goals and sequence them in alignment with development stage. Finally, alternate model types could be explored to analyze the interaction of input and model architecture; as the primary goal of this study was to evaluate the degree to which synthetic data can approximate naturalistic developmental input, this remains an open direction for future work. Overall, this work demonstrates the potential for prompt engineering in developmental language modeling, by abstracting over the details in a way that results in improved performance.

6 Limitations

Because our questions were generated by LLMs, we cannot guarantee their resemblance to questions in human CDS beyond quantitative measures such as MLU. Some of the questions generated by GPT-5-mini (Singh et al., 2025) diverge from topics and structures often observed in CDS, such as asking "were people hurt?" after a section about fatal accidents. This could also be an artifact of the topics included in the data, which are not fully child-appropriate, e.g., protions of the data extracted from The Open Subtitles corpus (Lison and Tiedemann, 2016).

In order to conserve resources and remain consistent with the BabyLM data, we restrict question generation to small-scale data. Future work could extend this approach to larger datasets, a greater number of training epochs, and additional model architectures. Crucially, while answers were generated for a subset of the question data, the current analysis is limited to questions alone. Whether the inclusion of answers yields additional gains in performance remains an open question.

Furthermore, we only test the GPT-Wee architecture (Bunzeck and Zarrieß, 2023), and therefore our results can only be interpreted within that scope.

Future work should extend this approach to additional model architectures.

Acknowledgments

We would like to extend our thanks to Anna Feldman and Jing Peng, members of the NLP Lab at Montclair State University, and reviewers for their support and feedback.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

BNC Consortium. 2007. The british national corpus, xml edition.

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

Bastian Bunzeck and Sina Zarrieß. 2023. GPT-wee: How small can a small language model really get? In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 35–46, Singapore. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Hu, Jaap Jumelet, Tal Linzen, Jing Liu, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Wilcox, and Adina Williams. 2025a. Babylm turns 3: Call for papers for the 2025 babylm workshop. *Preprint*, arXiv:2502.10645.

Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y. Hu, Jing Liu, Jaap Jumelet, Tal Linzen, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Gotlieb Wilcox, and Adina Williams. 2025b. Findings of the third BabyLM challenge: Accelerating language modeling research with cognitively plausible data. In *Proceedings of the First BabyLM Workshop*, pages 399–420, Suzhou, China. Association for Computational Linguistics.

Richard Diehl Martinez, Zébulon Goriely, Hope McGovern, Christopher Davis, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2023. CLIMB – curriculum learning for infant-inspired model building. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 112–127, Singapore. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *Preprint*, arXiv:2305.07759.

Steven Y. Feng, Noah D. Goodman, and Michael C. Frank. 2024. Is child-directed speech effective training data for language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22055–22071, Miami, Florida, USA. Association for Computational Linguistics.

Anita Gelboim and Elior Sulem. 2025. TafBERTa: Learning grammatical rules from small-scale language acquisition data in Hebrew. In *Proceedings of the First BabyLM Workshop*, pages 76–90, Suzhou, China. Association for Computational Linguistics.

Martin Gerlach and Francesc Font-Clos. 2020. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22(1):126.

Akari Haga, Akiyo Fukatsu, Miyu Oba, Arianna Bisazza, and Yohei Oseki. 2024. BabyLM challenge: Exploring the effect of variation sets on language model training efficiency. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 252–261, Miami, FL, USA. Association for Computational Linguistics.

Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21, Miami, FL, USA. 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.

Jaap Jumelet, Lucas Charpentier, Michael Hu, and Jing Liu. 2025. Babylm_2025. OSF.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2011. The Winograd schema challenge. In

AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning, volume 46, page 47.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.

Miyu Oba, Akari Haga, Akiyo Fukatsu, and Yohei Oseki. 2023. BabyLM challenge: Curriculum learning based on sentence complexity approximating language acquisition. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 290–297, Singapore. Association for Computational Linguistics.

Francesca Padovani, Bastian Bunzeck, Manar Ali, Omar Momen, Arianna Bisazza, Hendrik Buschmeier, and Sina Zarrieß. 2025a. Dialogue is not enough to make a communicative BabyLM (but neither is developmentally inspired reinforcement learning). In *Proceedings of the First BabyLM Workshop*, pages 421–435, Suzhou, China. Association for Computational Linguistics.

Francesca Padovani, Jaap Jumelet, Yevgen Matusevych, and Arianna Bisazza. 2025b. Child-directed language does not consistently boost syntax learning in language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19735–19756, Suzhou, China. Association for Computational Linguistics.

Whitney Poh, Michael Tombolini, and Libby Barak. 2025. What did you say? generating child-directed speech questions to train LLMs. In *Proceedings of the First BabyLM Workshop*, pages 237–245, Suzhou, China. Association for Computational Linguistics.

Jessica F Schwab and Casey Lew-Williams. 2016. Repetition across successive sentences facilitates young children's word learning. *Developmental psychology*, 52(6):879.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Nikitas Theodoropoulos, Giorgos Filandrianos, Vassilis Lyberatos, Maria Lymperaiou, and Giorgos Stamou. 2024. BERTtime stories: Investigating the role of synthetic story data in language pre-training. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 308–323, Miami, FL, USA. Association for Computational Linguistics.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Brian MacWhinney, and Chris Dyer. 2016. Learning the curriculum with Bayesian optimization for task-specific word representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 130–139, Berlin, Germany. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. *Preprint*, arXiv:1804.07461.

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. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Wikimedia. 2023. Simple english wikipedia dump. `https://dumps.wikimedia.org/simplewiki/20230301/`. Accessed: 2023-07-31.

Yue Yu, Elizabeth Bonawitz, and Patrick Shafto. 2019. Pedagogical questions in parent–child conversations. *Child development*, 90(1):147–161.

Author Index