ChineseDevBench: A Chinese Developmental Benchmark for Language Development

Shaonan Wang[1,*†], **Yiwen Wu**[2,*], **Na Li**[3,*], **Zesheng Chen**[3], **Gan Wang**[3], **Shuchen Zhang**[3], **Xin Sun**[1], **Luan Li**[4], **Yaran Chen**[3,†]

[1]Department of Language Science and Technology, Hong Kong Polytechnic University

[2]Faculty of Arts and Education, University of Auckland

[3]Xi'an Jiaotong-Liverpool University

[4]School of Foreign Languages, Shanghai Jiao Tong University

shaonan.wang@polyu.edu.hk, yaran.chen@xjtlu.edu

Abstract

How similar are the learning trajectories of language models and children? Recent work has narrowed the data-efficiency gap by training language models on child-scale input—roughly 10^8 tokens by early adolescence. However, evaluation remains largely based on adult-oriented English benchmarks and rarely involves direct comparison with human developmental data. We introduce ChineseDevBench, a Mandarin developmental benchmark comprising eight tasks that probe word meaning comprehension, association structure, acquisition dynamics, and language production. Crucially, the benchmark includes behavioral data from both children and adults, enabling direct model–human comparison. We train Chinese GPT-2 models on child-scale Mandarin input using an age-based curriculum and evaluate alignment between model and human response patterns across training checkpoints. Across most tasks, alignment improves with training but developmental age does not predict performance. ChineseDevBench provides a framework for systematically characterizing where model learning converges with—and diverges from—human language development.

Keywords: language development, learning trajectory, language models

1. Introduction

Humans acquire language rapidly with limited explicit supervision, progressing from isolated words to structured, meaning-rich utterances. Explaining this trajectory—and identifying which aspects are universal versus language-specific—remains a central goal of developmental research (Bowerman and Levinson, 2001).

Mandarin Chinese provides a particularly important test case. As the world's most widely spoken native language, Mandarin differs typologically from English in morphology, argument structure, and form–meaning mappings. Prior work shows that developmental trajectories in Mandarin syntax and grammatical–semantic marking diverge from English patterns (Huang et al., 2022; Ma et al., 2009). Corpus studies further document systematic age-related shifts in semantic expression beyond vocabulary growth (Tang et al., 2023). These findings motivate benchmarks that track Mandarin language development across multiple dimensions of meaning.

In parallel, modern language models achieve strong performance through web-scale training far exceeding human input. A child encounters on the order of 10^8 tokens by early adolescence, whereas many models are trained on orders of magnitude more. Initiatives such as BabyLM attempt to match child-scale exposure, but evaluation remains largely adult-centric and English-based, with limited comparison to child behavioral data (Warstadt et al., 2023; Hu et al., 2024; Charpentier et al., 2025; Tan et al., 2024). Moreover, recent work suggests a gap between linguistic competence and human-like processing. Steuer et al.

* Equal contribution.

† Corresponding author.

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 10–24
16 May 2026.

(2023) show that models achieving strong performance on linguistic benchmarks often exhibit a poorer fit to human reading times, pointing to a divergence between linguistic competence and psycholinguistic measures.

These findings point to two related concerns. First, it remains unclear whether results obtained from English-based models generalize cross-linguistically in the context of human language development. Second, it remains unclear whether this divergence between linguistic competence and human-like processing extends to other dimensions of language development, such as analogy and production.

To address these gaps, we introduce ChineseDevBench, a Mandarin developmental benchmark spanning eight tasks, including child and adult behavioral data probing lexical, semantic, and relational knowledge. Mandarin, with its distinct typological properties, offers an important test case for whether current findings generalize cross-linguistically. We train Chinese GPT-2 models on child-scale Mandarin input using a chronological curriculum and evaluate model–human alignment across training checkpoints, allowing us to track developmental trajectories and identify systematic convergences and divergences with human learning.

2. ChineseDevBench

ChineseDevBench consists of eight tasks organized into two groups: five child-oriented tasks (three unreleased and two publicly available) and three publicly available adult-oriented tasks. Together they probe compositional reasoning, associative structure, category knowledge, age of acquisition, production patterns, and perceptual feature structure.

2.1. Child-Oriented Tasks

CHILD-ANALOGY: Compositional Analogy

This task assesses morpho-semantic pattern completion through compositional analogy (The task originates from (Sun et al., 2022), and uses an unpublished dataset collected by co-author Sun). Children receive a short Mandarin prompt providing an example mapping from a description to a name, followed by a minimally changed description (15 in total). For example: "有一个又大又红的花叫大红花，那么有一个又大又黄的花叫什么? (There is a big, red flower called a big red flower; then what would you call a big, yellow flower?)"→ "大黄花 *(big yellow flower)*".

Population: 117 Mandarin-speaking children aged 3–6 years, divided into three age bins: 3–4 years (n=6), 4–5 years (n=63), and 5–6 years (n=48).

Evaluation: For each item, we compute the conditional log probability of the target word given the stimulus prompt (ending with 我们叫它 "we call it"). Performance is measured by Spearman correlation between children's accuracy and model scores.

CHILD-ASSOC: Word Association

This task evaluates cue–response association structure (574 in total) using child production data[1].

Population. 733 Mandarin-speaking children aged 2.5–7.5 years, divided into five age bins: 2.5–3.5 years (n=11), 3.5–4.5 years (n=102), 4.5–5.5 years (n=238), 5.5–6.5 years (n=306), and 6.5–7.5 years (n=76).

Evaluation. We use Representational Similarity Analysis (RSA). A human distance matrix is constructed from cue–response probability distributions; a corresponding matrix is derived from model embeddings. Alignment is measured via Spearman correlation between flattened matrices.

CHILD-FLUENCY: Verbal Fluency (Animals and Fruits)

This task assesses category (semantic) fluency. Children receive a category prompt (e.g., 请说出尽可能多的动物 "Please name as many animals as you can" / 请说出尽可能多的水果 "Please name as many fruits as you can") and produce an ordered list of category members. Data were collected from Li and Hills (2026), yielding 223 animal

[1] Unpublished data collected by co-author Li's group.

names and 90 fruit names.

Population: 733 Mandarin-speaking children aged 2.5–7.5 years, divided into five age bins consistent with CHILD-ASSOC.

Evaluation: For each category, we compute the conditional log probability of the target word given the stimulus prompt (e.g., 动物有 "Animals include" / 水果有 "Fruits include". Performance is measured by the Spearman correlation between the model's scores and children's production frequencies (occurrences divided by the number of children) at each age bin.

CHILD-WORD-ACQ: Word Acquisition

This task compares model-estimated age of acquisition (AoA) with human norms.

Source: Wordbank item data[2] and AOA data from Liu et al. (2011).

Population: Mandarin-speaking children from existing AoA norm datasets: 1,056 children aged 16–30 months from Wordbank (n≈70 per monthly age bin), and 262 children from Liu et al. (2011) across three kindergarten levels—K1 (n=99), K2 (n=64), and K3 (n=99).

Evaluation: Following (Chang and Bergen, 2022), we estimate model AoA by tracking token-level surprisal across training checkpoints. For each target word, we extract all sentences containing that word from adult speech in the CHILDES corpus and compute surprisal at each checkpoint. A sigmoid function is fitted to the resulting learning curve, and model AoA is operationalized as the training step at which surprisal crosses a threshold midway between baseline and minimum. Human–model alignment is assessed using Spearman rank correlation between model-predicted and child AoA. We selected 309 of 797 Wordbank words and 66 of 435 words from Liu et al. (2011), excluding items with insufficient occurrences for reliable estimation.

CHILD-PRODUCTION: Language Production

This task evaluates the model's ability to generate child-like continuations conditioned on short prefixes, and compares distributional properties of the generated output with developmental norms from child speech.

Source: CHILDES Mandarin corpus and CPCSLD (Chinese Preschool Children's Spoken Language Database)(Feng et al., 2026). Both corpora consist of transcribed child speech or lexical terms, and we use the existing transcriptions provided by these resources rather than raw audio.

Population: Mandarin-speaking children across developmental stages: <24, 24–30, 30–36, 36–42, 42–48, 48–60, 60–72, and 72+ months in CHILDES; 648 children from the CPCSLD dataset across three kindergarten levels—K1 (n=184, M=49.72 months, SD=4.08), K2 (n=181, M=60.60 months, SD=4.58), and K3 (n=283, M=71.36 months, SD=4.74).

Evaluation: We use short high-frequency prefixes selected from child utterances in the CHILDES Mandarin corpus (e.g., 我 喜 欢..."I like...") as prompts, and generate continuations using stochastic decoding (top-k = 50, top-p = 0.95, temperature = 0.9), with a maximum length of 50 tokens. For each checkpoint, 150 utterances are generated. To ensure linguistic validity and comparability with early child productions, we extract the first sentence segment (delimited by punctuation) from each generated continuation as the unit of analysis. The resulting utterances are segmented and part-of-speech tagged using jieba[3]. For each checkpoint, we compute lexical properties as well as shallow and sentence-level syntactic properties. These include the proportion of major part-of-speech categories (nouns, verbs, adjectives, adverbs), along with syntactic measures such as the proportion of clausal subjects. We then compare these distributions against benchmarks from two sources: lexical norms from CPCSLD (K1, K2, K3, ages 3–6), and distributional estimates computed from CHILDES child speech across eight age groups.

Note that the short prefixes used as generation prompts may also occur in the model's training data, as they correspond to high-frequency utterance-initial fragments commonly produced by

[2] https://wordbank.stanford.edu/data/?name=item_data

[3] https://github.com/fxsjy/jieba

young children. However, this does not constitute data contamination in the conventional sense. The prefixes are used solely as conditioning context, and the evaluation focuses on the distributional properties of the generated continuations rather than on the prefixes themselves. Moreover, such fragments are ubiquitous in child-produced speech, and their presence in training data is both unavoidable and ecologically appropriate.

2.2. Adult-Oriented Tasks

ADULT-STS: Semantic Textual Similarity

This task evaluates semantic textual similarity between sentence pairs.

Source: ChineseSTS[4] comprising 14,743 sentence pairs with human similarity ratings.

Evaluation: We compute Spearman's rank correlation between the model's cosine similarity scores and human gold-standard similarity scores for sentence pairs.

ADULT-ASSOC: SWOW-ZH Mandarin Word Association Norms

This task uses multiple-response free association (three responses per cue) with aggregate associative strengths.

Source: SWOW-ZH repository[5], providing word association data for 10,192 cues and over 2 million responses collected between 2016–2023 (Li et al., 2024).

Evaluation: RSA evaluation method identical to CHILD-ASSOC.

ADULT-SEMFEAT-6: Semantic Feature Norms

This task predicts semantic feature ratings across six dimensions: Vision, Motor, Socialness, Emotion, Time, and Space.

Source: Six Semantic Dimension Database[6], containing subjective ratings for 17,940 Chinese words (Wang et al., 2023).

Evaluation: For each target word, we extract its static representation by applying mean pooling over the last hidden layer. A linear regression probe maps these embeddings to human perceptual ratings using 5-fold cross-validation. Alignment is evaluated using Spearman's rank correlation between cross-validated model predictions and human ratings.

3. Language Model

3.1. Training Dataset

The training data comprise three categories: transcripts from children's cartoons, digitized children's picture books and audiobooks, and recordings of child-directed speech.

Children's Cartoons: We collected approximately 2,000 cartoon videos from Bilibili and YouTube, including roughly 1,000 videos targeting children aged 0–3 and 1,000 targeting children aged 3–6. Video selection was guided by age-specific keywords and titles, which were cross-referenced with children's media platforms[7] to ensure age appropriateness. Transcripts were generated using OpenAI's Whisper automatic speech recognition system.

Children's Picture Books and Audiobooks: We collected 1,158 digitized storybooks for children aged 0–6. Age labels and content were obtained from an open-access repository of Chinese children's literature that provides corresponding audio recordings for each story[8]. Audio content was transcribed using OpenAI's Whisper automatic speech recognition system.

CHILDES Mandarin Corpus: A core component of our conversational data was sourced from the TalkBank CHILDES Mandarin database[9]. We applied filtering criteria to select files with clear age markings and appropriate speaker roles. The selected data encompasses multiple

[4] https://github.com/IAdmireu/ChineseSTS

[5] https://github.com/lib314a/SWOWZH

[6] https://www.nature.com/articles/s41597-023-01995-6

[7] https://v.qq.com/channel/child

[8] https://www.limaogushi.com/

[9] https://talkbank.org/childes/access/Chinese/Mandarin/

sub-corpora including Beijing, Chang (Chang1, Chang2, ChangPlay), Erbaugh, LiReading, LiZhou, NSCtoys, Tong, and the comprehensive Zhou collection (Zhou1, Zhou2, Zhou3, ZhouAssessment, ZhouDinner, ZhouNarratives).

3.2. Training Settings

We implemented a decoder-only Transformer model based on the GPT-2 Small architecture, consisting of 12 Transformer blocks, 12 attention heads, and a hidden size of 768. We adopted the `bert-base-chinese` tokenizer (vocabulary size ≈21,128). Although originally derived from large-scale adult corpora, this tokenizer operates largely at the character level for Mandarin Chinese, where individual characters typically constitute stable and meaningful units and inflectional morphology is minimal. As a result, it approximates character-level modeling while maintaining compatibility with widely used Chinese NLP resources. The total parameter count is approximately 100M.

Training follows a chronological curriculum simulating child language acquisition, divided into two sequential stages:

Stage 1 (0–3 years): Focuses on simpler lexical items and sentence structures, comprising approximately 1.56 million tokens (across 5,126 extracted sentences). Training duration: 10 epochs (≈770 total steps). Checkpoints saved at each epoch end (11 checkpoints).

Stage 2 (3–6 years): Expands to approximately 2.97 million tokens (across 13,713 extracted sentences). Training duration: 10 epochs (≈5,700 total steps). Checkpoints saved at each epoch end (11 checkpoints).

Cumulative exposure totals approximately 4.53 million tokens. This checkpointing strategy enables reconstruction of the continuous learning trajectory across developmental phases.

3.3. Baseline Model

For comparative benchmarking of adult-level linguistic competence, we use UER GPT-2 Chinese[10]. This model shares identical architecture with our cognitive agent but was pre-trained on CLUECorpusSmall (approximately 14GB of general-domain Chinese text). This baseline quantifies the gap between models trained on limited, developmentally plausible child-directed speech and models trained on extensive open-domain data.

4. Results

We evaluated the developmental trajectory of our curriculum-trained model across ChineseDevBench. All model embeddings were extracted from the last layer. Results are organized into four sections: child-oriented comprehension and association tasks, word acquisition dynamics, language production, and adult-oriented semantic tasks.

4.1. Child-Oriented Tasks

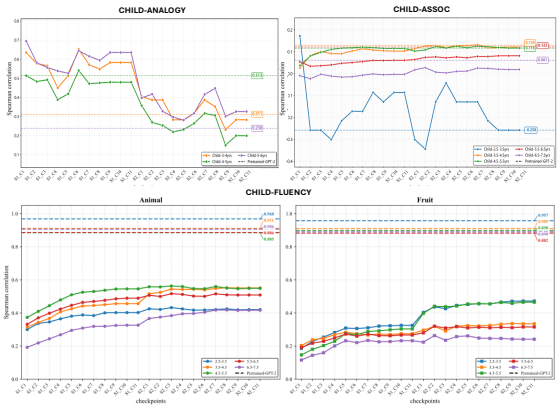

Fig. 1: GPT-2 developmental trends on CHILD-ANALOGY, CHILD-ASSOC, and CHILD-FLUENCY tasks.

Figure 1 presents results for three child-oriented tasks. For CHILD-ANALOGY, the first training stage achieves comparable or higher correlations with human responses than the pre-trained model, but performance drops to pre-trained levels during the second stage, suggesting that models trained on smaller datasets may

[10] https://huggingface.co/uer/gpt2-chinese-cluecorpussmall

better approximate child-like behavior for certain properties. For CHILD-ASSOC and CHILD-FLUENCY, model–human alignment generally improves with training, indicating that curriculum learning on child-directed input induces increasingly human-like semantic representations. For CHILD-ASSOC, RSA correlations remain low for both child-GPT and pretrained GPT-2, with greater variability in the 2.5–3.5 age bin due to smaller sample sizes; children's developmental age does not correspond to better model alignment, suggesting that alternative evaluation metrics may be needed. For CHILD-FLUENCY, a substantial performance gap exists between pretrained and child-GPT2 models, and developmental age bins show no consistent relationship with model performance.

4.2. Word Acquisition Dynamics (CHILD-WORD-ACQ)

We compared model and child word acquisition by tracking surprisal across training checkpoints and fitting sigmoid learning curves (Figure 2). Model AoA was defined as the training step at which surprisal crossed a threshold midway between baseline and minimum, analogous to the 50% production criterion in child studies.

Shared acquisition patterns. Both the model and children acquire certain word classes early: kinship terms (e.g., 奶奶 "grandmother," 哥哥 "older brother"), common animals (狗 "dog," 兔子 "rabbit"), basic household objects (电话 "phone"), and high-frequency everyday items (车 "car," 饭 "rice," 肉 "meat"). Conversely, both acquire abstract or low-imageability items late, such as the honorific classifier 位 and the plural pronoun 你们 "you-PL."

Divergent acquisition patterns. Two systematic dissociations emerged. First, grammatical function words (conjunctions, temporal markers) were acquired rapidly by the model but late by children (e.g., 已经 "already": model 142 tokens vs. child 29 months; 因为 "because": model 147 tokens vs. child 30 months). Second, embodied vocabulary—action verbs and sensory adjectives—showed the opposite pattern: children acquired these early (19–21 months), whereas the model required thousands of tokens (e.g., 带 "carry": model 4583 tokens vs. child 20 months; 香 "fragrant": model 2648 tokens vs. child 19 months).

These patterns suggest that the model efficiently exploits distributional regularities to learn function words, while children rely on conceptual development. Conversely, children leverage sensorimotor experience for action and sensory vocabulary, a grounding mechanism unavailable to text-only models. Detailed word lists are provided in Appendix Tables A1–A4.

Comparison with English. Consistent with Chang and Bergen (2022), we observed a negative effect of word length on model AoA: shorter words were acquired later in training. This contrasts with child language acquisition, where shorter words are typically learned earlier. Children are more likely to store short chunks of speech rather than chunks that are merely frequent or predictable (Grimm et al., 2019), a tendency shaped by phonological simplicity, memory constraints, and social interaction. Language models, in contrast, rely solely on distributional patterns in text, where shorter and more polysemous words are harder to represent consistently. Thus, the same surface property can have opposite effects depending on the underlying learning mechanism.

Unlike in English unidirectional models, where nouns tend to be acquired later than adjectives and verbs, our Chinese GPT-2 model acquired nouns relatively early. Conversely, verbs, acquired early by English models, were acquired later in the Chinese model. This pattern likely reflects several distributional challenges. Chinese verbs are often polysemous and context-dependent, making them more difficult to learn from distributional input alone. In contrast, Chinese nouns are more semantically stable, providing cleaner distributional signals. Additionally, the relatively flexible word order of Chinese allows verbs to appear in variable syntactic positions, potentially reducing the consis-

tency of their distributional contexts compared to nouns.

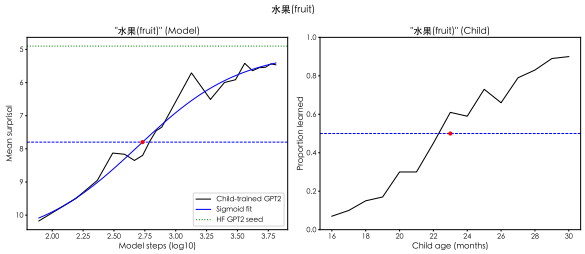

Fig. 2: Learning curves for the word "水果"(fruit) in a GPT-2 language model and human children. The blue horizontal lines indicate age of acquisition cutoffs. The green horizontal line reflects the baseline computed from the HF GPT-2 model prior to any child-directed training. The blue curve represents the fitted sigmoid function based on the language model surprisals during training (black). Child AoA values obtained from Hao et al. (2008).

4.3. Language Production (CHILD-PRODUCTION)

We examined whether GPT-2 exhibits child-like developmental trajectories in lexical complexity and syntactic category usage, comparing model outputs across checkpoints with CPCSLD benchmarks (kindergarten ages 3–6).

Lexical development. Table 1 and Figure 3 show that GPT-2 reproduces most qualitative developmental patterns: one-syllable words decrease while multi-syllable words increase, and noun proportions rise while adjective proportions fall. Five of six key indicators align with CPCSLD trajectories, providing partial support for human-like lexical development under curriculum learning.

Syntactic development. Figure 4 reveals a divergence between GPT-2 and children. Children in CHILDES exhibit low initial syntactic complexity followed by steady increases, reflecting staged acquisition of embedding structures. In contrast, GPT-2 shows substantial variability but no systematic growth, suggesting that the model accesses complex constructions early without undergoing human-like incremental syntactic development.

Table 1: Production trends in word length in syllables and POS distribution: GPT-2 checkpoints (Model) vs. CPCSLD benchmarks (Ref). *Align* indicates directional consistency with reference trajectories.

Feature	Model	Δ%	Ref	Align
Word Length				
1-syllable	↓	-7.63	↓	Yes
2-syllable	↑	+4.34	↑	Yes
3+ syllable	↑	+2.83	↑	Yes
POS Types				
Noun	↑	+3.97	↑	Yes
Verb	↑	+3.40	↓	No
Adjective	↓	-3.79	↓	Yes

4.4. Adult-Oriented Tasks

Figure 5 presents results for three adult-oriented tasks. Consistent with the child task results, a performance gap exists between pre-trained GPT-2 and child-GPT2 models, and model–human alignment generally improves with training for ADULT-STS and ADULT-SEMFEAT-6. For ADULT-STS and ADULT-ASSOC, performance drops sharply from the first to second training stage before recovering, suggesting that these tasks are sensitive to shifts in embedding distributions. Similar to CHILD-ASSOC, ADULT-ASSOC yields low correlation scores overall. For ADULT-SEMFEAT-6, the six semantic dimensions show differential alignment with human ratings: Vision, Socialness, and Space achieve higher correlations, while Time and Emotion prove most difficult to align—possibly because affective and temporal semantics require experiential grounding beyond what distributional statistics can provide.

5. Discussion

Human language acquisition has been explained by competing theoretical frameworks. Nativist accounts propose that children rely on innate linguistic constraints (Chomsky, 2002), while usage-based approaches argue that language emerges from actual use (Tomasello, 2003; Ib-

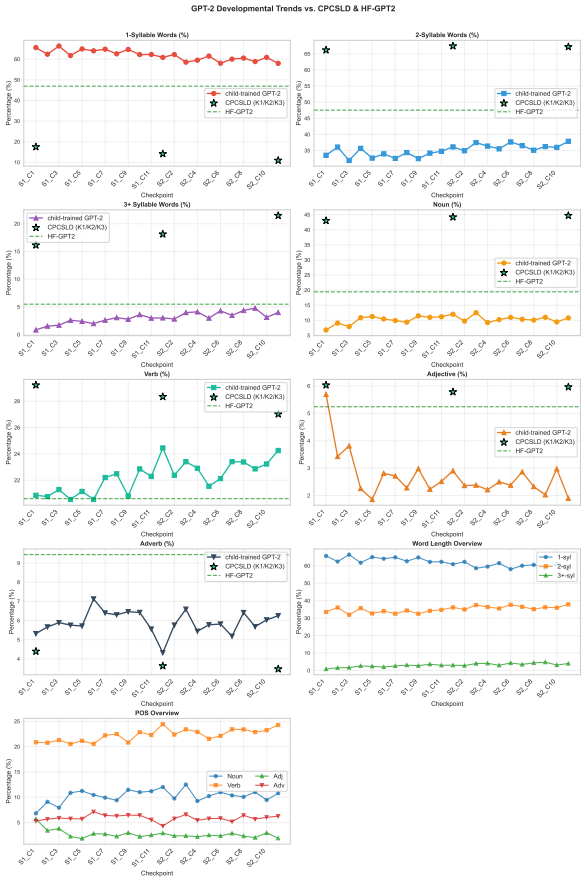

Fig. 3: Production: GPT-2 developmental trends compared with CPCSLD, focusing on lexical measures.

botson, 2013). Usage-based theories emphasize sensitivity to recurring patterns in input. Language models are likewise sensitive to such regularities in linguistic data. However, models lack the social-cognitive dimension that usage-based theorists consider equally essential to language development.

Following Portelance and Jasbi (2024), we interpret the model's behavior as jointly informative about the scope and limits of distributional learning. The model's strengths indicate aspects of language that can be captured from distributional information, while its limitations point to domains in which additional mechanisms—sensorimotor, social, or cognitive—likely play a crucial role. These mechanisms do not uniformly accelerate learning, but may shape different aspects of acquisition in distinct ways. The model's delayed acquisition of action verbs and sensory adjectives,

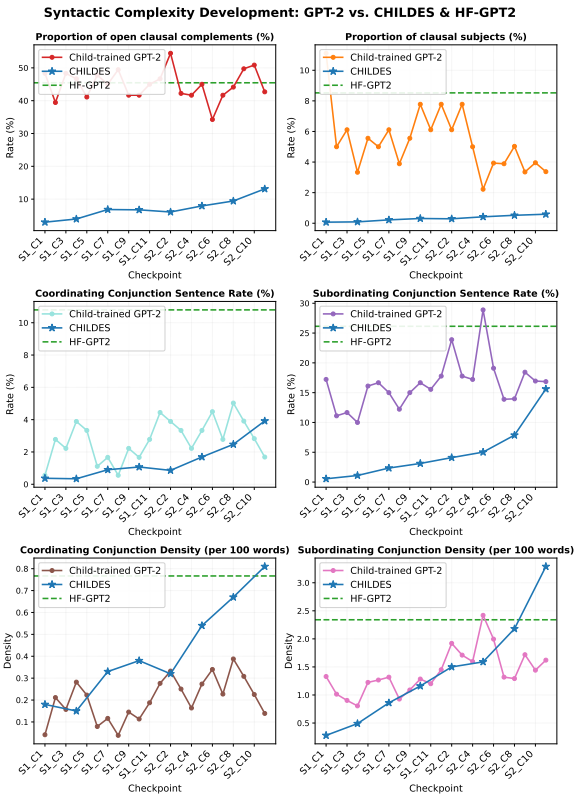

Fig. 4: Production: GPT-2 developmental trends relative to CHILDES-derived syntactic measures. Blue stars denote syntactic measures computed from CHILDES child speech for eight age groups (<24, 24–30, 30–36, 36–42, 42–48, 48–60, 60–72, and 72+ months), with one point per age group.

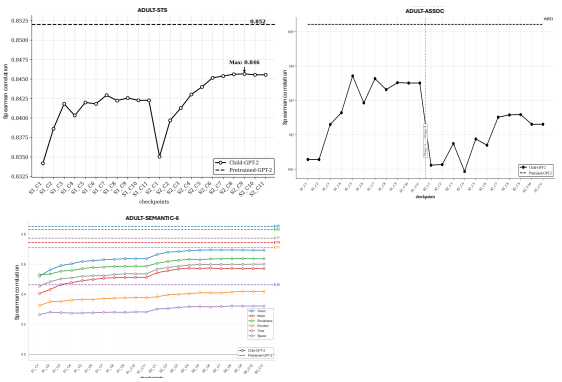

Fig. 5: GPT-2 developmental trends on ADULT-STS, ADULT-ASSOC and SEMFEAT-6 tasks.

which are typically acquired early in child development, implicates that sensorimotor grounding constitutes an important inductive bias in children's language acquisition. Conversely, children's comparatively late acquisition of function words, de-

spite their high distributional regularity in child-directed speech, suggests that their acquisition may depend on conceptual and pragmatic capacities that develop gradually and are not solely determined by input frequency. Taken together, these divergences suggest that human learners and language models may rely on different underlying mechanisms, even when models achieve similar outcomes.

Future work may explore how such mechanisms can be incorporated into computational models, for example through multimodal input or interactive learning settings, bringing models of language acquisition closer to the cognitive processes they aim to approximate.

6. Conclusion

ChineseDevBench provides a Mandarin developmental benchmark integrating child and adult behavioral data across eight tasks. By training GPT-2 on child-scale input with a chronological curriculum, we track model learning trajectories and compare them directly with human development. Results reveal both convergence and systematic divergence. For most child-oriented tasks, alignment improves with training, yet developmental age does not predict performance and syntactic complexity lacks human-like growth; the model rapidly acquires function words but lags on embodied vocabulary. For adult-oriented tasks, alignment also improves with training, though association tasks yield low correlations, and dimensions like Time and Emotion remain difficult to align, suggesting they require experiential grounding. Unlike English models, which acquire nouns later than adjectives and verbs, the Chinese GPT-2 model acquired nouns early and verbs later, highlighting language-specific effects on acquisition order. ChineseDevBench enables more precise evaluation of developmentally plausible language models and provides a foundation for future work on grounding, curriculum design, and cross-linguistic developmental comparison. *The data and evaluation code are publicly available at* *https://github.com/Evaisgreat/Chinese-DevBench*.

7. Ethical statements and limitations

All child data are drawn from publicly available datasets, datasets that will be made publicly available, or previously published datasets collected under appropriate ethical oversight. No new child data were collected for this study, and no personally identifiable information is included. The sampled populations may not fully represent the diversity of Mandarin-speaking children.

Part of the training data were automatically transcribed using OpenAI's Whisper speech recognition system, which may introduce transcription errors. Some evaluation metrics rely on specific prompting methods, and results may vary under different prompts. In addition, the available child data are relatively limited in size; more publicly released datasets will be necessary to obtain more robust and generalizable findings. Model–child comparisons are behavioral rather than mechanistic, and observed alignment does not imply shared underlying cognitive processes. The benchmark focuses on lexical and semantic development and does not comprehensively cover pragmatic or discourse-level abilities. Finally, the benchmark is intended for research purposes only and should not be used for clinical assessment.

8. Acknowledgements

This work was supported by the start-up fund project (1-BDE3) sponsored by the Faculty of Humanities of the Hong Kong Polytechnic University, the Research Development Fund (RDF) of Xi' an Jiaotong-Liverpool University (XJTLU) under Grant No. RDF-24-02-028 and Suzhou Science and Technology Development Planning Programme (Grant No. ZXL2025310）.

9. References

Melissa Bowerman and Stephen C Levinson. 2001. *Language acquisition and conceptual development*. 3. Cambridge University Press.

Tyler A Chang and Benjamin K Bergen. 2022. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16.

Lucas Georges Gabriel Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y Hu, Jing Liu, Jaap Jumelet, Tal Linzen, Aaron Mueller, Candance Ross, et al. 2025. Findings of the third babylm challenge: Accelerating language modeling research with cognitively plausible data. In *Proceedings of the First BabyLM Workshop*, pages 399–420.

Noam Chomsky. 2002. *Syntactic structures*. Walter de Gruyter.

Chen Feng, Shuo Wang, and Shuang Li. 2026. Cpcsld: A lexical database of chinese preschool children's spoken words. *Behavior Research Methods*, 58(2):54.

Robert Grimm, Giovanni Cassani, Steven Gillis, and Walter Daelemans. 2019. Children probably store short rather than frequent or predictable chunks: Quantitative evidence from a corpus study. *Frontiers in psychology*, 10:80.

Min Hao, Hua Shu, Aiqing Xing, and Ping Li. 2008. Early vocabulary inventory for mandarin chinese. *Behavior Research Methods*, 40(3):728–733.

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. *arXiv preprint arXiv:2412.05149*.

Ruoyu Lexie Huang, Paul Fletcher, Zhixiang Zhang, Weilan Liang, Virginia Marchman, and Twila Tardif. 2022. Early grammatical marking development in mandarin-speaking toddlers. *Developmental psychology*, 58(4):631.

Paul Ibbotson. 2013. The scope of usage-based theory. *Frontiers in psychology*, 4:255.

Bing Li, Ziyi Ding, Simon De Deyne, and Qing Cai. 2024. A large-scale database of mandarin chinese word associations from the small world of words project. *Behavior Research Methods*, 57(1):34.

Luan Li and Thomas T. Hills. 2026. Exploration-versus-exploitation of semantic networks promotes early lexical development: Evidence from preschoolers' semantic fluency and computational simulations (unpublished pre-print). *Retrieved from https://osf.io/7t3hn/*.

Youyi Liu, Meiling Hao, Ping Li, and Hua Shu. 2011. Timed picture naming norms for mandarin chinese. *PloS one*, 6(1):e16505.

Weiyi Ma, Roberta Michnick Golinkoff, Kathy Hirsh-Pasek, Colleen McDonough, and Twila Tardif. 2009. Imageability predicts the age of acquisition of verbs in chinese children. *Journal of child language*, 36(2):405–423.

Eva Portelance and Masoud Jasbi. 2024. The roles of neural networks in language acquisition. *Language and Linguistics Compass*, 18(6):e70001.

Julius Steuer, Marius Mosbach, and Dietrich Klakow. 2023. Large gpt-like models are bad babies: A closer look at the relationship between linguistic competence and psycholinguistic measures. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 142–157.

Xin Sun, Kehui Zhang, Rebecca A Marks, Nia Nickerson, Rachel L Eggleston, Chi-Lin Yu, Tai-Li Chou, Twila Tardif, and Ioulia Kovelman. 2022. What's in a word? cross-linguistic influences on spanish–english and chinese–english bilingual children's word reading development. *Child development*, 93(1):84–100.

Alvin Tan, Chunhua Yu, Bria Long, Wanjing Ma, Tonya Murray, Rebecca Silverman, Jason Yeatman, and Michael C Frank. 2024. Devbench: A multimodal developmental benchmark for language learning. *Advances in Neural Information Processing Systems*, 37:77445–77467.

Tempo Po-Yi Tang, Dustin Kai-Yan Lau, and Man-Tak Leung. 2023. Corpus of mandarin child language: a preliminary study on the acquisition of semantic content categories in mandarin-speaking preschoolers. *Frontiers in psychology*, 14:1234525.

Michael Tomasello. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Shaonan Wang, Yunhao Zhang, Weiting Shi, Guangyao Zhang, Jiajun Zhang, Nan Lin, and Chengqing Zong. 2023. A large dataset of semantic ratings and its computational extension. *Scientific Data*, 10(1):106.

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*, pages 1–34.

Appendix: Word Acquisition Detailed Results

Table 2: Twenty words with the earliest Age of Acquisition. Comparison between GPT-2 model learning steps and children's AoA (months), based on Wordbank data. Words in **bold** indicate overlap between the model and children.

2*Rank	Model Earliest				Child Earliest			
	Word	Model AoA	Child AoA	Type	Word	Child AoA	Model AoA	Type
1	**车 (car)**	**81.15**	**16**	**vehicles**	**车 (car)**	**16**	**81.15**	**vehicles**
2	火车 (train)	95.75	21	vehicles	**阿姨 (auntie)**	**16**	**103.45**	**people**
3	**阿姨 (auntie)**	**103.45**	**16**	**people**	**奶奶 (grandma)**	**16**	**122.23**	**people**
4	今天 (today)	115.67	23	time_words	哥哥 (brother)	16	146.94	people
5	告诉 (tell)	119.24	23	action_words	球 (ball)	16	353.36	toys
6	**奶奶 (grandma)**	**122.23**	**16**	**people**	抱 (hug)	16	379.43	action_words
7	喜欢 (like)	124.81	21	action/descriptive	妈妈 (mom)	16	693.57	people
8	小朋友 (kids)	125.87	20	people	爸爸 (dad)	16	962.13	people
9	桌子 (table)	126.12	19	furniture_rooms	拿 (take)	16	994.86	action_words
10	现在 (now)	126.27	23	time_words	一 (one)	16	1275.79	quantifiers
11	朋友 (friend)	126.32	22	people	灯 (light)	16	1348.66	household
12	张 (classifier)	127.61	23	classifiers	**叔叔 (uncle)**	**17**	**141.80**	**people**
13	干嘛 (what for)	129.94	21	question_words	蛋 (egg)	17	381.34	food_drink
14	长 (grow/long)	129.98	21	action_words	脚 (foot)	17	161.11	body_parts
15	漂亮 (pretty)	131.85	22	descriptive_words	笔 (pen)	17	170.87	supplies
16	就 (then)	135.40	23	connecting_words	饭 (rice)	17	174.70	food_drink
17	已经 (already)	141.73	29	time_words	肉 (meat)	17	327.37	food_drink
18	**叔叔 (uncle)**	**141.80**	**17**	**people**	飞 (fly)	17	333.03	action_words
19	飞机 (airplane)	144.21	20	vehicles	书 (book)	17	344.84	supplies
20	嘛 (particle)	144.99	21	final_particles	水 (water)	17	360.68	food_drink

Table 3: Twenty words with the latest Age of Acquisition. Comparison between GPT-2 model learning steps and children's AoA (months), based on Wordbank data. Words in **bold** indicate overlap between the model and children.

12*Rank	Model Latest				Child Latest			
	Word	Model AoA	Child AoA	Type	Word	Child AoA	Model AoA	Type
1	带 (bring/carry)	4582.65	20	action_words	**位 (classifier)**	**30**	**1572.94**	**classifiers**
2	藏 (hide)	3687.71	20	action_words	得 (must/particle)	30	278.38	helping_verbs
3	路 (road)	3068.54	21	outside	因为 (because)	30	146.92	connecting_words
4	写 (write)	2987.26	20	action_words	已经 (already)	29	141.73	time_words
5	香 (fragrant)	2648.09	19	descriptive_words	以后 (after/later)	29	388.75	time_words
6	啦 (particle)	2413.04	22	final_particles	时间 (time)	29	184.85	time_words
7	小 (small)	2189.56	19	descriptive_words	这些 (these)	29	372.13	pronouns
8	**梯子 (ladder)**	**2093.66**	**24**	**outside**	他们 (they)	28	341.20	pronouns
9	舔 (lick)	2052.29	21	action_words	不错 (not bad)	27	492.72	descriptive_words
10	阳台 (balcony)	1764.43	23	furniture_rooms	全部 (all)	27	681.46	quantifiers
11	拉 (pull)	1711.31	20	action_words	些 (some)	27	422.81	classifiers
12	腿 (leg)	1588.03	19	body_parts	别的 (other)	27	688.07	pronouns
13	**位 (classifier)**	**1572.94**	**30**	**classifiers**	向 (toward)	26	563.22	directions
14	哭 (cry)	1462.79	20	action_words	晚 (late/night)	25	723.19	time_words
15	等 (wait)	1442.76	20	action_words	我们 (we)	25	514.47	pronouns
16	**你们 (you-plural)**	**1423.67**	**25**	**pronouns**	**你们 (you-plural)**	**25**	**1423.67**	**pronouns**
17	踢 (kick)	1406.86	19	action_words	人家 (others)	25	269.66	pronouns
18	开心 (happy)	1383.12	22	descriptive_words	次 (times/classifier)	25	315.10	classifiers
19	滚 (roll)	1346.09	22	action_words	层 (floor/layer)	25	963.55	classifiers
20	龙 (dragon)	1312.99	19	action_words	**梯子 (ladder)**	**24**	**2093.66**	**outside**

Table 4: Twenty words with the earliest Age of Acquisition. Comparison between GPT-2 model learning steps and children's AoA (years), based on Liu et al. (2011). Words in **bold** indicate overlap between the model and children.

2*Rank	Model Earliest				Child Earliest			
	Word	Model AoA	Child AoA	Type	Word	Child AoA	Model AoA	Type
1	**狗 (dog)**	**-0.53**	**2.48**	**animals**	**兔子 (rabbit)**	**1.94**	**65.83**	**animals**
2	火 (fire)	17.73	5.19	natural	**眼睛 (eyes)**	**1.94**	**98.46**	**body**
3	牛 (cow)	44.10	7.71	animals	蝴蝶 (butterfly)	1.94	346.22	animals
4	杯子 (cup)	46.86	4.44	supplies	**飞机 (airplane)**	**2.25**	**67.60**	**vehicles**
5	鼻子 (nose)	49.39	3.74	body	苹果 (apple)	2.25	123.63	food
6	桌子 (table)	62.55	2.85	furniture	门 (door)	2.25	179.40	furniture
7	熊 (bear)	65.26	10.45	animals	**房子 (house)**	**2.30**	**69.67**	**buildings**
8	**兔子 (rabbit)**	**65.83**	**1.94**	**animals**	**火车 (train)**	**2.30**	**75.23**	**vehicles**
9	**飞机 (airplane)**	**67.60**	**2.25**	**vehicles**	**狗 (dog)**	**2.48**	**-0.53**	**animals**
10	**房子 (house)**	**69.67**	**2.30**	**buildings**	碗 (bowl)	2.48	159.83	supplies
11	盒子 (box)	71.96	5.68	supplies	**电话 (phone)**	**2.54**	**80.20**	**electronics**
12	**火车 (train)**	**75.23**	**2.30**	**vehicles**	**太阳 (sun)**	**2.54**	**88.28**	**natural**
13	**电话 (phone)**	**80.20**	**2.54**	**electronics**	乌龟 (turtle)	2.54	287.21	animals
14	头发 (hair)	80.91	10.40	body	长颈鹿 (giraffe)	2.63	321.48	animals
15	狮子 (lion)	85.37	3.78	animals	糖 (sugar)	2.63	1049.10	food
16	**太阳 (sun)**	**88.28**	**2.54**	**natural**	书 (book)	2.84	155.34	supplies
17	**眼睛 (eyes)**	**98.46**	**1.94**	**body**	床 (bed)	2.84	403.65	furniture
18	电视 (TV)	100.37	4.18	electronics	桌子 (table)	2.85	62.55	furniture
19	**老鼠 (mouse)**	**117.79**	**2.85**	**animals**	**老鼠 (mouse)**	**2.85**	**117.79**	**animals**
20	沙发 (sofa)	122.94	3.42	furniture	葡萄 (grape)	2.86	208.09	food

Table 5: Twenty words with the latest Age of Acquisition. Comparison between GPT-2 model learning steps and children's AoA (years), based on Liu et al. (2011). Words in **bold** indicate overlap between the model and children.

12*Rank	Model Latest				Child Latest			
	Word	Model AoA	Child AoA	Type	Word	Child AoA	Model AoA	Type
1	手臂 (arm)	1094.99	3.56	body	信 (letter)	10.68	357.11	abstract
2	糖 (sugar)	1049.10	2.63	food	**鹿 (deer)**	**10.64**	**275.26**	**animals**
3	心 (heart)	571.16	4.58	body	**苍蝇 (fly)**	**10.45**	**410.90**	**animals**
4	**针 (needle)**	**554.09**	**7.44**	**tools**	熊 (bear)	10.45	65.26	animals
5	腿 (leg)	462.44	3.41	body	头发 (hair)	10.40	80.91	body
6	斑马 (zebra)	429.76	4.12	animals	屋顶 (roof)	7.84	343.40	buildings
7	**苍蝇 (fly)**	**410.90**	**10.45**	**animals**	牛 (cow)	7.71	44.10	animals
8	床 (bed)	403.65	2.84	furniture	狼 (wolf)	7.45	247.23	animals
9	蚂蚁 (ant)	372.23	4.58	animals	**针 (needle)**	**7.44**	**554.09**	**tools**
10	猪 (pig)	365.25	3.68	animals	盒子 (box)	5.68	71.96	supplies
11	信 (letter)	357.11	10.68	abstract	尾巴 (tail)	5.19	155.75	body
12	**船 (boat)**	**355.66**	**4.20**	**vehicles**	火 (fire)	5.19	17.73	natural
13	蝴蝶 (butterfly)	346.22	1.94	animals	窗户 (window)	4.73	176.12	furniture
14	屋顶 (roof)	343.40	7.84	buildings	心 (heart)	4.58	571.16	body
15	龙 (dragon)	333.07	3.56	animals	蚂蚁 (ant)	4.58	372.23	animals
16	长颈鹿 (giraffe)	321.48	2.63	animals	礼物 (gift)	4.56	202.00	supplies
17	桥 (bridge)	313.87	3.73	buildings	杯子 (cup)	4.44	46.86	supplies
18	乌龟 (turtle)	287.21	2.54	animals	医生 (doctor)	4.42	253.32	people
19	**鹿 (deer)**	**275.26**	**10.64**	**animals**	**船 (boat)**	**4.20**	**355.66**	**vehicles**
20	**钟 (clock)**	**270.72**	**4.20**	**supplies**	**钟 (clock)**	**4.20**	**270.72**	**supplies**