# HSKBenchmark: Modeling and Benchmarking Chinese Second Language Acquisition in Large Language Models through Curriculum Tuning

**Qihao Yang[1*], Xuelin Wang[2*], Jiale Chen[1], Xuelian Dong[1], Yuxin Hao[3†], Tianyong Hao[1†]**

[1]School of Computer Science, South China Normal University, Guangzhou, China
[2]College of Chinese Language and Culture, Jinan University, Guangzhou, China
[3]School of Chinese Studies and Exchange, Shanghai International Studies University, Shanghai, China
{charlesyeung, jlchen, xldong, haoty}@m.scnu.edu.cn, wangxuelin@stu2022.jnu.edu.cn, hyx_tcfl@163.com

## Abstract

Language acquisition is vital to revealing the nature of human language intelligence and has recently emerged as a promising perspective for improving the interpretability of large language models (LLMs). However, it is ethically and practically infeasible to conduct experiments that require controlling human learners' language inputs. This poses challenges for the verifiability and scalability of language acquisition modeling, particularly in Chinese second language acquisition (SLA). While LLMs provide a controllable and reproducible alternative, a systematic benchmark to support phase-wise modeling and assessment is still lacking. To address these issues, we propose HSKBenchmark, the first benchmark for staged modeling and writing assessment of LLMs in Chinese SLA. The benchmark covers HSK levels 3 to 6, comprising authentic textbooks with 6.76M tokens, 16K synthetic instruction data, 30 test topics and a linguistically-grounded evaluation system. To simulate human acquisition trajectories, a curriculum-tuning framework is introduced, which trains LLMs in a progression from beginner to advanced proficiency levels. Since language production in writing is a key perspective for observing SLA development, an evaluation system is established to probe LLMs in writing, including the coverage of level-based grammar items, writing errors, lexical complexity, syntactic complexity, and holistic scoring. We also develop an HSKAgent fine-tuned on 10K compositions from Chinese second language learners to automate this evaluation system. Extensive experimental results demonstrate that HSKBenchmark not only models Chinese SLA effectively, but also serves as a reliable benchmark for dynamic writing assessment in LLMs. Our fine-tuned LLMs have writing performance on par with advanced human learners and exhibit human-like acquisition characteristics. The HSKBenchmark and HSKAgent serve as foundational tools and resources, with the potential to pave the way for future research on language acquisition modeling and LLMs interpretability.

**Code** — https://github.com/CharlesYang030/HSKB
**Datasets** — https://github.com/CharlesYang030/HSKB
**Extended version** — https://arxiv.org/abs/2511.15574

[*]These authors contributed equally.
[†]Corresponding authors.



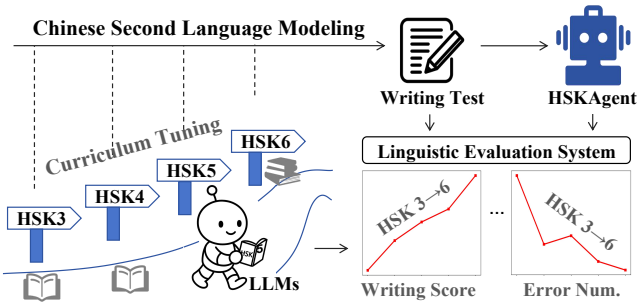


Figure 1: Illustration of Chinese SLA modeling and dynamic writing assessment in LLMs.

## Introduction

Since the mid-20th century, research on language acquisition has advanced rapidly, laying theoretical foundations for understanding human language intelligence (Chomsky 1965; Lenneberg 1967; Chomsky 1980). However, due to ethical and practical limitations, many experiments involving controlled language inputs and the simulation of learning trajectories are difficult to conduct with human learners (Warstadt and Bowman 2022). As a result, the field has long faced challenges in terms of verifiability and computational modeling. Against this backdrop, large language models (LLMs) emerge as a valuable resource because of their controllability and reproducibility. The language acquisition of LLMs is receiving increasing attention. Researchers suggest that modeling the developmental patterns in LLMs not only enhances interpretability but also provides new theoretical and empirical insights into human language learning mechanisms (Warstadt and Bowman 2020).

Language acquisition is mainly categorized into first language (L1) acquisition and second language (L2) acquisition (SLA). Existing studies have explored L1 acquisition modeling of language models by adjusting the neural network architecture, optimizing hyperparameter settings, introducing linguistic features, or applying causal intervention (Warstadt and Bowman 2022). They achieve success in simulating children's vocabulary and grammar acquisition. Researchers attempt to transfer such success to SLA modeling. For example, a recent work trains XLM (Conneau and Lample 2019) from scratch using a L1-L2 parallel corpus and ob-

serves that the model has similarities to humans in the transfer pattern from L1 to L2 (Oba et al. 2023). However, the SLA modeling of LLMs remains unresolved due to the lack of level-based training data and evaluation systems. Existing methods (Aoyama and Schneider 2024a) simply limit the size of training data rather than considering the difficulty of acquiring L2, resulting in unclear boundaries in SLA stages. Although different multilingual benchmarks are widely used to probe LLMs on various multilingual tasks, they mainly evaluate LLMs' existing capabilities (Hendrycks et al. 2021; Ahuja et al. 2024) rather than dynamic assessment for SLA modeling. Importantly, there are approximately 375 million English L2 learners and 20 million Chinese L2 learners in the world. The huge group stimulates an urgent need for empirical research on SLA modeling.

This paper studies an important yet overlooked issue: SLA modeling and dynamic writing assessment in LLMs, as shown in Figure 1. The **applicability of LLMs** is first considered: modeling SLA in LLMs requires selecting a non-English target language as L2, since most language models are trained primarily on large-scale English data. The **data accessibility** is also considered: there are extensive learning materials in Chinese, as *Hanyu Shuiping Kaoshi* (HSK) (Peng, Yan, and Cheng 2021) is a representative Chinese L2 proficiency test. The **assessment method** is further considered: language production in writing is a key perspective for observing L2 development (Durrant, Brenchley, and McCallum 2021), which has advantages of reflecting the mastery of LLMs in the use of language structures. Based on these three considerations, in order to provide a reusable evaluation framework for SLA modeling, a feasible solution is to build a benchmark from the perspective of Chinese as L2 to assess the language output in writing of LLMs. Importantly, Chinese is an isolating language typologically distinct from English (Huang 2015). Studying Chinese SLA modeling can be a representative view to examine whether LLMs can generalize across typologically diverse languages and capture structural patterns beyond Indo-European norms.

However, to achieve this goal, we encounter three major challenges. The first challenge is to build a benchmark with level-based training data. This requires using training data with clear level boundaries to distinguish acquisition stages developmentally, rather than merely controlling the scale of training data as in existing studies (Aoyama and Schneider 2024a; Constantinescu et al. 2025). The second challenge is to simulate human-like staged acquisition in LLMs and track its progression. This requires a curriculum-based design that incrementally exposes LLMs to staged Chinese inputs. The third challenge is to create an efficient evaluation system. This requires integrating linguistically-grounded indicators for LLMs writing and automating the system.

To address these challenges, we propose **HSKBenchmark**, the first benchmark for staged modeling and writing assessment of LLMs in Chinese SLA. To construct level-based training data, we collect 79 widely-used textbooks in international Chinese education, covering HSK levels 3 to 6. These textbooks with 6.76M tokens are used for staged pretraining. Following the *Chinese Proficiency Grading Standards for International Chinese Language Education*, we identify 591 grammar items annotated with HSK levels. Three state-of-the-art LLMs (GPT, DeepSeek, Gemini) with robust Chinese capabilities are prompted to generate instruction data for writing exercises based on these grammar items. The 16k generated data is used for staged fine-tuning, with an agreement score of 0.91 and a validity rate of 95%. In addition, thirty writing topics from real HSK exams are set as testing tasks. To simulate human-like staged acquisition, we introduce a curriculum-tuning framework, enabling LLMs to undergo staged pretraining followed by instruction tuning at each stage from HSK levels 3 to 6. For assessment, we build an evaluation system grounded in five linguistic dimensions: the coverage of level-based grammar items, writing errors, lexical complexity, syntactic complexity, and holistic scoring. We further develop an HSKAgent, an automated evaluator fine-tuned on the grammar dataset and 10K compositions from human Chinese L2 learners.

Our main contributions are summarized as follows:

- The HSKBenchmark is proposed, which is the first benchmark for staged modeling and writing assessment of LLMs in Chinese SLA. It has the potential to serve as foundational tools and resources for future research on language acquisition modeling.
- A curriculum-tuning framework is introduced to simulate human language acquisition trajectories, and an HSKAgent is also developed to automate our linguistically-grounded evaluation system.
- Extensive experiments demonstrate the effectiveness of HSKBenchmark. Our fine-tuned LLMs achieve high writing performance on par with advanced human learners, contributing to the verification of SLA theories.

## Related Work

### Language Acquisition Modeling with Neural Language Models

There has been much debate about the mechanism of language acquisition for a long time (Warstadt and Bowman 2022). To investigate the nature of language acquisition, neural language models were employed for language acquisition modeling in the 1980s (Rumelhart and McClelland 1985; Pinker and Prince 1988). Although these early models had limited linguistic capabilities, their integration with cognitive science provided experimental insights into language mechanisms. In the past decade, with the advancement of natural language processing technology, language acquisition modeling has received renewed attention (Warstadt and Bowman 2022). While Dupre (Dupre 2021) points out that language models lack real language learning capabilities, an increasing number of researchers believe they can be utilized as effective tools to verify language acquisition theories (Warstadt and Bowman 2022; Futrell and Mahowald 2025).

Existing work focuses mainly on modeling L1 acquisition (Warstadt and Bowman 2022) to investigate the difference of inductive bias between human and machine (McCoy, Frank, and Linzen 2020; Warstadt et al. 2020). A recent work uses inductive bias distillation to transfer the

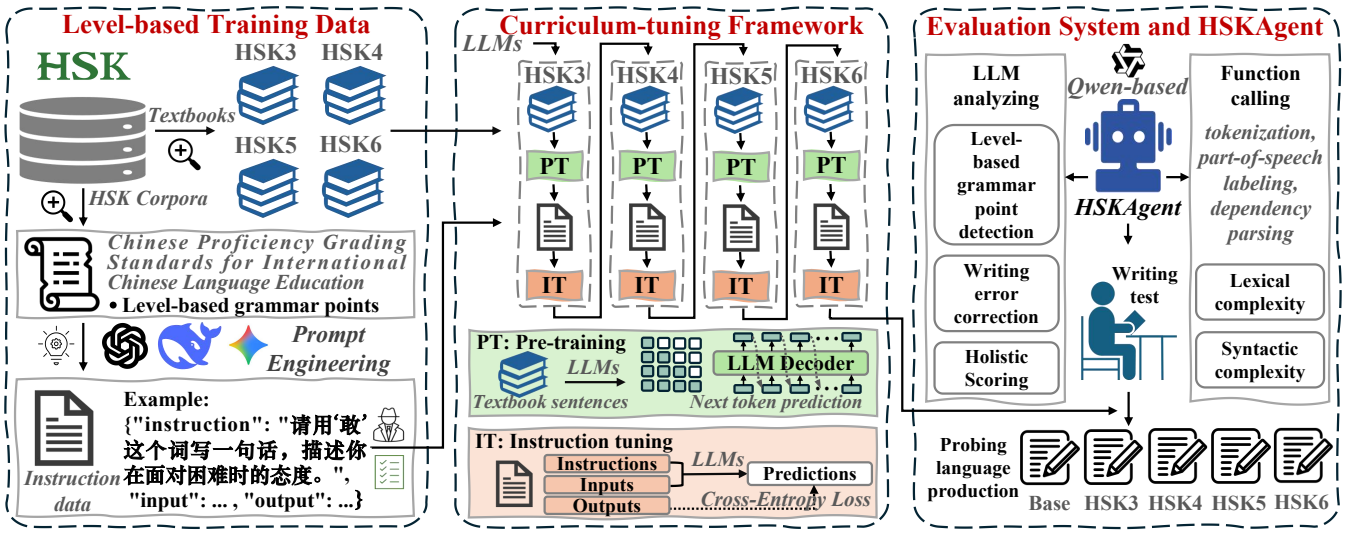


Figure 2: Illustration of our HSKBenchmark. It contains the level-based training data, the curriculum-tuning framework, the linguistically-grounded evaluation system and the HSKAgent.

Bayesian priors into the neural network (McCoy and Griffiths 2025). The research shows that such models not only learn languages from limited data, but also acquire complicated syntactic structures from large-scale corpora. Besides, many studies manipulate the internal structure of the models through controlling neural architectures (Yedetore et al. 2023) and hyperparameters (Chang and Bergen 2022), or explore the structural bias of the models using linguistic features (Ravfogel et al. 2020) or causal interventions (Finlayson et al. 2021). A shared task named BabyLM (Warstadt et al. 2023; Hu et al. 2024) was proposed recently to promote the development of evaluation frameworks for modeling child language acquisition.

In contrast, research on SLA modeling is still at an early stage and focuses primarily on L1–L2 transfer (Warstadt and Bowman 2022; Aoyama and Schneider 2024b). A recent study explores the effects of L1-L2 transfer in the XLM model across different L1 (French, German, Russian, Japanese) and English as L2, finding that L1 pretraining significantly enhanced L2 syntactic generalization (Oba et al. 2023). The results indicate that transfer effects are influenced by typological distances and training configuration. However, such studies roughly distinguish the stages of language acquisition by controlling corpus size, lacking systematic modeling of the developmental trajectory of L2 production, especially in the context of Chinese as a second language (Aoyama and Schneider 2024a; Constantinescu et al. 2025). Therefore, this paper aims to adopt a curriculum-based approach and investigate the development of LLMs in linguistic competence in writing during the process of Chinese SLA modeling.

## Resources and Evaluation in Chinese SLA

The *Hanyu Shuiping Kaoshi* (HSK) is currently the most widely-used standardized test to assess the Chinese proficiency of non-native Chinese learners (Peng, Yan, and Cheng 2021). It consists of six levels (1 to 6) like Common European Framework of Reference for language (CEFR) (Council of Europe 2001), and provides a comprehensive evaluation of language skills including listening, speaking, reading, and writing. Many teaching resources are organized according to HSK levels, such as *Developing Chinese* and *Chinese Course*. In addition, open-access learner corpora like the *HSK Dynamic Composition Corpus* contain manually annotated error corrections and proficiency scores. These materials offer a diverse and level-based source of training data for our work.

Linguistic complexity indices are widely used to evaluate the writing performance of Chinese L2 learners (Hao et al. 2024; Hao, Wang, and Lin 2022; Hao et al. 2023). The CTAP for Chinese (Cui et al. 2022) achieves the automated extraction of 196 linguistic complexity indices across character, word, sentence, and paragraphs for Chinese learner writing. However, it does not calculate writing scores, which are a key indicator for measuring SLA development. While L2C-Rater (Wang and Hu 2021) predicts essay scores through regression models that integrate linguistic features, pre-extracted writing errors, and textual features, it lacks the ability to automatically detect errors for new compositions. Moreover, scoring essays through human teachers incurs high costs and low efficiency. Therefore, this paper aims to incorporate linguistic indicators that are specifically relevant to Chinese SLA development into the evaluation system, and to leverage LLMs with robust Chinese capabilities to develop an efficient agent for automated scoring.

## The HSKBenchmark

To propose the HSKBenchmark, we make efforts from the construction of the level-based training data, the design of a curriculum-tuning framework, the development of a linguistically-grounded evaluation system and an HSKAgent, as shown in Figure 2.

| Textbook levels | Tokens | Sentences | Average number of tokens per sentence |
|---|---|---|---|
| HSK 3 | 895,037 | 22,743 | 39.35 |
| HSK 4 | 1,473,516 | 34,637 | 42.66 |
| HSK 5 | 1,717,178 | 41,044 | 41.84 |
| HSK 6 | 2,678,621 | 63,650 | 42.08 |
| Total | 6,764,352 | 162,074 | 41.74 |

Table 1: Statistics of the level-based textbooks.

| Items | HSK3 | HSK4 | HSK5 | HSK6 | Advan. | Total |
|---|---|---|---|---|---|---|
| Word | 110 | 48 | 47 | 50 | 62 | 317 |
| Phrase | 9 | 6 | 8 | 11 | 21 | 55 |
| FF | 5 | 6 | 6 | 3 | 5 | 25 |
| SC | 14 | 4 | 11 | 3 | 7 | 39 |
| ST | 27 | 27 | 26 | 16 | 47 | 143 |
| EU | 3 | 4 | 3 | 1 | 1 | 12 |
| ALL | 168 | 95 | 101 | 84 | 143 | 591 |
| Num. | 4,600 | 2,607 | 2,896 | 2,334 | 4,025 | 16,462 |

Table 2: Statistics of the grammar items and the instruction data. Advan. refers to the advanced HSK level.

## The Construction of the Level-based Training Data

Krashen, one of the representative researchers in SLA research, argues that language acquisition occurs when learners are exposed incrementally to comprehensible input that contains linguistic features slightly beyond their current level (i+1) (Krashen 1982). In real L2 teaching scenarios, learners are also taught from beginner to advanced levels of teaching materials. However, existing studies do not pay attention to this issue because they usually distinguish the different stages of language acquisition based on the size of training data (Liu et al. 2024b; Aoyama and Schneider 2024a). For example, five learning stages can be divided in training data with 1 million tokens, where each batch of 200K tokens is regarded as one stage. In addition, the training data includes learning materials of different difficulties, without clearly distinguishing between beginner and advanced levels. To bridge this gap, we refer to the HSK level standard[1] that divides Chinese L2 proficiency into 6 levels, of which HSK levels 3 to 6 have writing tasks. Simultaneously, we conduct a survey of available resources for Chinese SLA. Two major issues are identified: (1) fewer learning materials are available at lower levels, particularly for HSK levels 1 and 2; (2) a substantial amount of manual effort is required to align multiple-choice questions in official HSK test collections with their corresponding answers, making it difficult to incorporate such items into the evaluation system like benchmarks in other domains.

To construct the level-based training data, we first collect 79 widely-used textbooks based on HSK levels 3 to 6, such as *HSK Standard Course* and *Boya Chinese Course*. These textbooks are a mixture of texts and images. We delete the images since multimodal inputs are not the objective of this study. In order to ensure the semantic compactness of the texts, we also delete all the Pinyin and English symbols used to assist learning in the textbooks through scripts. Finally, the total number of tokens in the cleaned textbooks is 6.76M, with 162,074 sentences and an average of 41.74 tokens per sentence, as shown in Table 1.

Besides textbooks, human teachers often ask learners to complete writing exercises to improve their language production ability. Therefore, we create a set of instruction data covering various writing exercises. Specifically, we first integrate HSK levels 3 to 6 and advanced grammar items from *Chinese Proficiency Grading Standards for International Chinese Language Education*. Six types of grammar items are selected because they appear at these levels, including word, phrase, fixed format (FF), sentence component (SC), sentence type (ST), and emphatic usage (EU). Data that include multiple words or usages in the same grammar item are manually split. Secondly, we leverage GPT-4.1-mini (Achiam et al. 2023), DeepSeek-Chat-V3 (Liu et al. 2024a), and Gemini-2.5-Flash (Team et al. 2023) with robust Chinese capabilities to generate level-based instruction data according to these grammar items using in-context learning with two shots. The LLMs are prompted to generate 10 instruction instances for each grammar item. Each piece of generated data contains an instruction, an input, and an output (as shown in the example in Figure 2), where the instruction is the requirement of a writing exercise, the input is the specified grammar item, and the output is the expected language production. Then, three graduate annotators are recruited and trained on HSK standards. A randomly sampled set from the generated data is manually verified by the annotators using Fleiss's Kappa, yielding an agreement score of 0.91 and a validity rate of 95%. Finally, we conduct proofreading and data filtering and then obtain 16,462 synthetic instruction data based on these 591 level-based grammar items. The statistics of the grammar items and the synthetic level-based instruction data are reported in Table 2.

## The Curriculum-tuning Framework

After distinguishing the stages in Chinese SLA using the level-based textbooks and instruction data, LLMs are also required to adapt to such staged modeling and assessment rather than being trained on all data at once. To this end, we introduce a curriculum-tuning framework, enabling LLMs to simulate Chinese L2 learners from self-learning on textbooks to writing exercises at each stage for gaining progressive capabilities in writing.

First, **pretraining on level-based textbooks for simulating input-based learning**: we define an HSK level $l \in \{3, 4, 5, 6\}$, and the corresponding level-specific textbooks are denoted as $\mathcal{T}^{(l)} = \{x_1, x_2, \ldots, x_m\}$, where each $x_i$ is a

[1]Chinese learners in HSK level 3 can use Chinese to complete basic communication tasks in life, study, work, etc. Those in HSK level 6 can easily understand the Chinese information heard or read, and express their opinions fluently in Chinese in oral or written form. Detailed level-by-level descriptions can be accessed at: https://www.chinesetest.cn/userfiles/file/dagang/HSK-koushi.pdf.

Chinese sentence. A LLM adopts a causal language modeling architecture and is trained using next-token prediction to compute the loss for each sentence. The pretraining loss at level $l$ is defined as:

$$\mathcal{L}_{\mathrm{PT}}^{(l)} = -\sum_{i=1}^{m}\sum_{t=1}^{|x_i|}\log P_{\theta^{(0)}}\left(x_{i,t} \mid x_{i<t}\right) \tag{1}$$

where $\theta^{(0)}$ denotes the LLM's initial parameters. For each sentence, $x_{i,t}$ refers to its $t$-th token, and $x_{i<t}$ denotes the preceding context before that token. After this stage of training, the resulting model is denoted as $LLM - \theta_{\mathrm{PT}}^{(l)}$.

Second, **instruction tuning on writing exercises for simulating output-based learning**: we use the instruction data $\mathcal{D}^{(l)} = \{(p_1, y_1), (p_2, y_2), \ldots, (p_n, y_n)\}$ corresponding to HSK level $l$, where each $p_i$ is a writing prompt and $y_i$ is the target completion. In this paper, the writing prompt is the combination of the instruction (the requirement of writing exercises) and the input (the specific grammar item), and the completion is the output (the expected language production). The LLM is then fine-tuned on this instruction-following task using the same language modeling loss:

$$\mathcal{L}_{\mathrm{IT}}^{(l)} = -\sum_{i=1}^{n}\sum_{t=1}^{|y_i|}\log P_{\theta_{\mathrm{PT}}^{(l)}}\left(y_{i,t} \mid p_i, y_{i<t}\right) \tag{2}$$

The resulting model after this stage of instruction tuning is denoted as $LLM - \theta_{\mathrm{IT}}^{(l)}$.

Finally, **curriculum tuning across levels**: LLMs experience curriculum tuning in ascending order of levels, namely from HSK level 3 to 6. At each level $l$, the LLM is first pretrained on the textbook data $\mathcal{T}^{(l)}$ and then instruction-tuned on the corresponding instruction data $\mathcal{D}^{(l)}$. The model parameters are updated at each level according to:

$$\theta_{\mathrm{PT}}^{(l)} = \mathrm{Pretraining}\left(\theta^{(l-1)}, \mathcal{T}^{(l)}\right) \tag{3}$$

$$\theta_{\mathrm{IT}}^{(l)} = \mathrm{InstructionTuning}\left(\theta_{\mathrm{PT}}^{(l)}, \mathcal{D}^{(l)}\right) \tag{4}$$

The final model $LLM - \theta^{(6)}$ is obtained by sequential fine-tuning on all level-based textbooks and instruction data, thereby simulating a complete Chinese SLA trajectory.

## The Linguistically-grounded Evaluation System and HSKAgent

To fairly evaluate the writing performance of LLMs, we collect 30 writing topics from the *HSK Dynamic Composition Corpus v2.0*[2] as test tasks. This corpus, released by Beijing Language and Culture University, is a collection of written compositions produced by non-native Chinese speakers from 85 countries (32.85% from Korea) in HSK test from 1992 to 2005. It includes more than 10K compositions with 4 million Chinese characters. These selected 30 topics cover a range of genres (e.g., narrative and argumentative writing) and topics (e.g., daily life and study). After examination, there is no data overlap or contamination between these 30 topics and our training data.

To capture and reflect the development of Chinese SLA across levels, we design an evaluation system by following previous work, covering five linguistic dimensions. (1)***The Coverage of Grammar Items*** refers to the proportion of grammar items from each HSK level in compositions. This metric is used to evaluate LLMs' mastery of grammar items across different proficiency levels. (2)***Writing Errors*** (**Err**) (Yan and Lin 2023) refers to the sum of character-level errors, lexical errors, syntactic errors and discourse-level errors. This metric is used to evaluate the accuracy of LLMs' language output. (3)***Lexical Complexity*** (**MATTR-50**) (Kyle et al. 2024) refers to the ratio of word types to word tokens within text windows, where each batch of 50 tokens is set as one window. This metric is used to evaluate LLMs' lexical proficiency. (4)***Syntatic Complexity*** (**MDD**) (Liu 2008) refers to the average dependency distance of texts. This metric is used to evaluate LLMs' syntactic proficiency. A higher MDD indicates longer dependency relations, which may reflect more sophisticated sentence structures. (5)***Holistic Scoring*** (**Score**) (Ramesh and Sanampudi 2022) refers to the overall score, which is typically determined based on the length, quality and the relevance of the text.

To automate the evaluation system, we develop an **HSK-Agent** built upon Qwen3-8B (Bai et al. 2023). The Qwen3-8B model is selected due to its strong performance in Chinese among 7/8B-scale models based on the SuperCLUE leaderboard[3]. It also has advantages in reproducibility and inference efficiency. Specifically, we transform the level-based instruction data into a binary classification dataset. For the positive samples, the original prompt and completion are concatenated into a new positive prompt, with the corresponding answer "Yes". For the negative samples, the prompt is paired with a negative completion randomly sampled from the data pool, resulting in a new negative prompt with the answer "No". To reduce the likelihood that the negative completion still aligns with the target grammar item, we restrict sampling to completions outside the current grammar item category. Although this is a straightforward approach, a manual validation yields an inter-annotator agreement score of 0.93 and a validity rate of 96%. Then, we reconstruct the original human-written versions from these 10K compositions with error annotations and scores. This dataset is used to train and test the HSK-Agent. Eventually, our HSKAgent achieves an F1-score of 0.97 for binary classification of grammar items, 90% accuracy for error detection, and an F1-score of 0.81 for holistic scoring. It also obtains good agreements with human raters (Quadratic Weighted Kappa (QWK) = 0.7969, Spearman = 0.8010, Pearson = 0.8023). For complexity-related indices, the HSKAgent leverages function calling for automatic computation.

# Experiments and Results

## Implementation Details

**Baselines.** Since our objective is not to train LLMs to acquire Chinese from scratch, we select LLMs that already

[2]https://yuyanziyuan.blcu.edu.cn/info/1043/1501.htm

[3]https://www.superclueai.com/

| Human/LLMs | The Coverage of Grammar Items | | | | | Writing Errors | Lexical Complexity | Syntactic Complexity | Holistic Scoring |
|---|---|---|---|---|---|---|---|---|---|
| | HSK3 | HSK4 | HSK5 | HSK6 | Advan. | Err | MATTR-50 | MDD | Score |
| Natives | 0.3408 | 0.2439 | 0.1745 | 0.1261 | 0.1146 | 1.4000 | 0.8061 | 2.9769 | 88.3333 |
| Leaner-95* | 0.3563 | 0.2040 | 0.1656 | 0.1392 | 0.1350 | **2.8667** | **0.8165** | **2.8386** | **85.0000** |
| Leaner-90* | 0.3481 | 0.1854 | 0.1997 | 0.1425 | 0.1243 | **3.3667** | **0.8059** | **2.9705** | **84.0000** |
| Leaner-80* | 0.3855 | 0.1914 | 0.1835 | 0.1327 | 0.1069 | **3.5000** | **0.7925** | **2.6473** | **74.8333** |
| Leaner-70* | 0.3802 | 0.2094 | 0.1978 | 0.1211 | 0.0915 | **3.8333** | **0.7764** | **2.6205** | **70.1667** |
| Leaner-60* | 0.3947 | 0.2030 | 0.1967 | 0.1034 | 0.1021 | **4.8000** | **0.7806** | **2.5814** | **63.0000** |
| GPT-4.1-mini | 0.3979 | 0.2324 | 0.1622 | 0.1082 | 0.0993 | 0.0000 | 0.8287 | 2.6032 | 91.5000 |
| DeepSeek-Chat | 0.4102 | 0.2118 | 0.1615 | 0.1166 | 0.0999 | 0.0000 | 0.8427 | 2.5411 | 92.3333 |
| Gemini-2.5 | 0.4038 | 0.2265 | 0.1673 | 0.1103 | 0.0921 | 0.0000 | 0.8334 | 2.5894 | 90.5300 |
| Llama2 | 0.4844 | 0.1615 | 0.1667 | 0.1126 | 0.0748 | 0.9000 | 0.6860 | 2.4253 | 70.0000 |
| Llama2$_{\text{HSK3}}$ | **0.4925**↑ | 0.1738 | 0.1471 | 0.1143 | **0.0723**↓ | **0.6333**↓ | **0.7188**↑ | **2.5045**↑ | **75.8333**↑ |
| Llama2$_{\text{HSK4}}$ | 0.4517 | **0.2048**↑ | 0.1768 | 0.0880 | **0.0787**↑ | **0.6667**↓ | **0.7364**↑ | **2.5503**↑ | **78.6667**↑ |
| Llama2$_{\text{HSK5}}$ | 0.4203 | 0.2005 | **0.1852**↑ | 0.1111 | **0.0829**↑ | **0.5667**↓ | **0.7592**↑ | **2.5274**↓ | **80.6667**↑ |
| Llama2$_{\text{HSK6}}$ | 0.4246 | 0.1818 | 0.1775 | **0.1279**↑ | **0.0883**↑ | **0.5333**↓ | **0.7641**↑ | **2.5558**↑ | **81.8333**↑ |
| Ch-Alpaca | 0.4470 | 0.2000 | 0.1678 | 0.1191 | 0.0661 | 0.0667 | 0.7705 | 2.5251 | 77.5000 |
| Ch-Alpaca$_{\text{HSK3}}$ | **0.4270**↓ | 0.1917 | 0.1803 | 0.1105 | **0.0905**↑ | **0.5000**↓ | **0.7774**↑ | **2.5329**↑ | **75.8333**↓ |
| Ch-Alpaca$_{\text{HSK4}}$ | 0.4049 | **0.2252**↑ | 0.1639 | 0.1069 | **0.0990**↑ | **0.0333**↓ | **0.7726**↓ | **2.5109**↓ | **82.0000**↑ |
| Ch-Alpaca$_{\text{HSK5}}$ | 0.3980 | 0.1859 | **0.2146**↑ | 0.1250 | **0.0765**↓ | **0.1000**↓ | **0.7816**↑ | **2.5557**↑ | **87.6667**↑ |
| Ch-Alpaca$_{\text{HSK6}}$ | 0.3844 | 0.2382 | 0.1632 | **0.1161**↓ | **0.0981**↑ | **0.0000**↓ | **0.7829**↑ | **2.5729**↑ | **85.6667**↑ |
| Mistral | 0.4798 | 0.1836 | 0.1603 | 0.1037 | 0.0726 | 0.7333 | 0.5260 | 2.5302 | 76.8333 |
| Mistral$_{\text{HSK3}}$ | **0.4542**↓ | 0.1802 | 0.1637 | 0.1190 | **0.0829**↑ | **0.5667**↓ | **0.7566**↑ | **2.5334**↑ | **79.5000**↑ |
| Mistral$_{\text{HSK4}}$ | 0.4006 | **0.2393**↑ | 0.1583 | 0.1141 | **0.0876**↑ | **0.4667**↓ | **0.7788**↑ | **2.5858**↑ | **81.1667**↑ |
| Mistral$_{\text{HSK5}}$ | 0.4020 | 0.1983 | **0.1719**↑ | 0.1222 | **0.1056**↑ | **0.3667**↓ | **0.7901**↑ | **2.5595**↓ | **82.3333**↑ |
| Mistral$_{\text{HSK6}}$ | 0.4141 | 0.1981 | 0.1437 | **0.1422**↑ | **0.1019**↓ | **0.3000**↓ | **0.7886**↓ | **2.6772**↑ | **85.3333**↑ |

Table 3: The Chinese SLA performance of human and LLMs on HSKBenchmark. Learners-*X** refers to those who got a original score of *X* in the *HSK Dynamic Composition Corpus v2.0*. Ch-Alpaca indicates the Chinese-Alpaca model. The upward and downward arrows indicate whether the model's current performance has improved or declined compared to its previous level.

possess a certain degree of Chinese capabilities, to investigate their developments during the Chinese SLA modeling. Therefore, we refer to SuperCLUE and choose three models of relatively low rank as baselines, including LLaMA2-7B-Chat, Mistral-7B-Instruct-v0.3, and Chinese-Alpaca-2-7B. Three stronger LLMs, GPT-4.0-mini, DeepSeek-Chat-V3, and Gemini-2.5-Flash, are also selected as baselines. Moreover, we include Chinese native speakers and Chinese L2 learners as human baselines.

**Setting.** The experiments are implemented on PyTorch 2.6.0 and 3 RTX 3090 GPUs (24GB) using LLaMA-Factory (Zheng et al. 2024). LoRA (Hu et al. 2022) is utilized to fine-tune these LLMs and the HSKAgent in pretraining and instruction tuning, where the learning rate is 5e-5, the number of epoch is 3 and bf16 is used as the compute type.

## Main Results

The Chinese SLA performance of human and LLMs on HSKBenchmark is reported in Table 3. Compared with Chinese SLA learners, the natives achieve the highest overall score (88.3333). Although the advanced learners (95* and 90*) also score more than 80, there is still a noticeable gap between them and the natives in terms of writing errors and syntactic complexity. Moreover, as learners improve their proficiency from 60* to 95*, their scores also gradually increase, which provides evidence that there is indeed a predictable developmental progress in Chinese SLA and our HSKAgent indeed presents such a trend reasonably. GPT, DeepSeek, and Gemini obtain average scores exceeding 90, but they are inferior to humans in syntactic complexity and mastery of advanced grammar items.

LLaMA2, Chinese-Alpaca, and Mistral all exhibit substantial improvements after Chinese SLA modeling. The base LLaMA2 model achieves a score of only 70, roughly equivalent to that of Learners-70*. After modeling at HSK3, LLaMA2$_{\text{HSK3}}$ improves by 5.83 points, and the final LLaMA2$_{\text{HSK6}}$ achieves a score of 81.83 on par with Learners-90*. In addition, the coverages of HSK3 and HSK4 grammars of LLaMA2$_{\text{HSK3}}$ are 49.25% and 17.38%, but LLaMA2$_{\text{HSK4}}$ shows a 4.08% decrease and a 3.10% increase respectively in these two aspects. This indicates that the curriculum-tuning framework enables the model to better acquire more complex grammars. Compared with LLMs$_{\text{HSK3}\sim\text{HSK4}}$, LLMs$_{\text{HSK5}\sim\text{HSK6}}$ get a higher proportion of advanced grammar items that are not included in training data. This suggests that more advanced models may develop emergent abilities to master higher-level grammars and generalize beyond the training data, much like the hu-

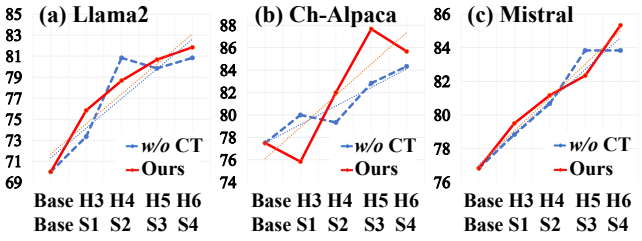


Figure 3: Comparison between our LLMs and those trained on the shuffled dataset. CT refers to the curriculum tuning. H*X* indicates HSK level *X* and S*X* indicates stage *X*.

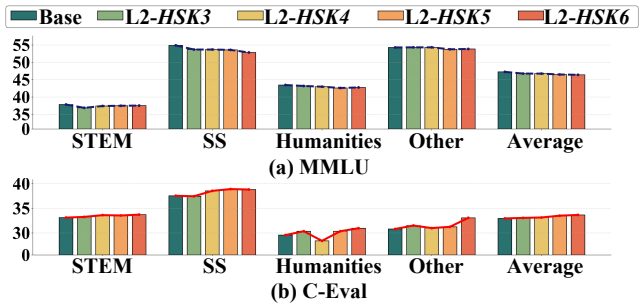


Figure 4: The performance of the curriculum-tuned Llama2 on MMLU and C-Eval. L2 indicates the Llama2 model and SS refers to Social Science.

man capacity to infer and extend learned knowledge.

Compared with human learners, LLMs are less prone to produce errors. A possible reason is that the language production mechanisms in writing of humans and LLMs are fundamentally different. Compared with LLMs, human writers might tend to take more risks in those usages they are not fully confident in. Limited by the top-k next-token prediction mechanism, LLMs tend to generate only those tokens in which they have the highest confidence. However, LLMs fall short of humans in lexical and syntactic complexity. LLMs optimize for predictive likelihood, tending to generate shorter, more typical sentences found in natural corpora. In contrast, L2 learners often deliberately use complex structures in writing tests to display linguistic competence, leading to higher syntactic complexity.

In summary, Table 3 presents comparisons between native speakers and L2 learners, between humans and LLMs, as well as the developmental trajectories of baseline LLMs in Chinese SLA. These results are consistent with expectations and support the effectiveness of our HSKBenchmark as an effective suite for benchmarking Chinese SLA performance.

## Ablation Study

An ablation study is conducted to reveal the effectiveness of our curriculum-tuning framework. Specifically, we shuffle and merge all level-based textbooks and instruction data into a single dataset. It is then divided into four stages (corresponding to HSK levels 3 to 6) purely based on data volume. The LLMs are finetuned on this dataset without level-based ordering. Figure 3 illustrates the comparison between our LLMs trained on the curriculum-tuning framework and those trained on the shuffled pretraining method in overall average scores. The results show that the shuffled approach enables LLMs to achieve relatively higher average scores in the early stages, likely because the models are exposed to high-level training data prematurely. However, in the later stages (stage 3-4), the performance of our LLMs surpasses that of the shuffled approach. This suggests that even when trained on the same data, an appropriate learning sequence is essential for activating better Chinese SLA outcomes in LLMs. This finding not only validates the effectiveness of our curriculum-tuning framework, but also aligns with Krashen's i+1 input hypothesis (Krashen 1982). This is because that our HSKBenchmark provides the training data with progressive difficulties like the i+1 input hypothesis which emphasizes the importance of progressively structured input in successful L2 acquisition.

## Impact on L1 Proficiency and General Chinese Performance

An additional experiment is conducted to examine whether LLMs' L1 proficiency and general Chinese abilities change during Chinese SLA modeling. Llama2 is selected to be evaluated on two benchmarks, MMLU (Hendrycks et al. 2021) and C-Eval (Huang et al. 2023). MMLU is a widely-used multitask English benchmark with QAs in STEM, social science (SS), humanities and other subjects. C-Eval is a widely-used comprehensive Chinese exam benchmark with similar QAs. The results, as shown in Figure 4, show that Llama2 does not suffer degradation in L1 performance (no catastrophic forgetting) on MMLU and even exhibits slight L2 improvements on C-Eval. This pattern is similar to the behavior of human L2 learners, showing that the curriculum-tuned LLMs trained on HSKBenchmark present human-like characteristics. This finding might support extending our method to other language frameworks like CEFR to uncover more empirical insights about SLA modeling.

## Conclusion

This paper proposes HSKBenchmark for staged modeling and writing assessment of LLMs in Chinese SLA. A curriculum-tuning framework is introduced to simulate human language acquisition trajectories. A linguistically-grounded evaluation system is designed to assess the language production of LLMs in writing, and an HSKAgent is developed to automate the evaluation system. Experimental results demonstrate that HSKBenchmark effectively supports Chinese SLA modeling in LLMs. The curriculum-tuning framework facilitates more robust SLA development compared to traditional training approaches, and the evaluation system and HSKAgent successfully capture and reflect this developmental progress. The suite of models developed in this work is released to serve as effective tools and resources for the community. In future work, we will scale the SLA modeling framework to a broader range of languages, incorporate multimodal inputs, and integrate additional linguistic dimensions to further explore the potential of LLMs in computational modeling and advancing SLA theories.

## Acknowledgements

The work was supported by grants from National Natural Science Foundation of China (No. 62372189), the Research Grants Council of the Hong Kong Special Administrative Region, China (UGC/FDS16/E03/25), Fujian Provincial Social Science Foundation Project (No. FJ2025B104) and "the Fundamental Research Funds for the Central Universities", and the China Scholarship Council (No. 202406780045).

## References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Ahuja, S.; Aggarwal, D.; Gumma, V.; Watts, I.; Sathe, A.; Ochieng, M.; Hada, R.; Jain, P.; Ahmed, M.; Bali, K.; and Sitaram, S. 2024. MEGAVERSE: Benchmarking Large Language Models Across Languages, Modalities, Models and Tasks. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2598–2637. Mexico City, Mexico: Association for Computational Linguistics.

Aoyama, T.; and Schneider, N. 2024a. Modeling Nonnative Sentence Processing with L2 Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 4927–4940. Miami, Florida, USA: Association for Computational Linguistics.

Aoyama, T.; and Schneider, N. 2024b. Modeling nonnative sentence processing with L2 language models. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, 4927–4940.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; et al. 2023. Qwen Technical Report. arXiv:2309.16609.

Chang, T. A.; and Bergen, B. K. 2022. Word Acquisition in Neural Language Models. *Transactions of the Association for Computational Linguistics*, 10: 1–16.

Chomsky, N. 1965. Aspects of the Theory of Syntax.

Chomsky, N. 1980. A Review of BF Skinner's Verbal Behavior. *The Language and Thought Series*, 48–64.

Conneau, A.; and Lample, G. 2019. Cross-lingual Language Model Pretraining. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Constantinescu, I.; Pimentel, T.; Cotterell, R.; and Warstadt, A. 2025. Investigating Critical Period Effects in Language Acquisition through Neural Language Models. *Transactions of the Association for Computational Linguistics*, 13: 96–120.

Council of Europe, X. 2001. *Common European framework of reference for languages: Learning, Teaching, Assessment*. Cambridge University Press.

Cui, Y.; Zhu, J.; Yang, L.; Fang, X.; Chen, X.; Wang, Y.; and Yang, E. 2022. CTAP for Chinese: A Linguistic Complexity Feature Automatic Calculation Platform. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 5525–5538.

Dupre, G. 2021. Can Deep Learning Contribute to Theoretical Linguistics? *Minds and Machines*, 31(4): 617–635.

Durrant, P.; Brenchley, M.; and McCallum, L. 2021. *Understanding Development and Proficiency in Writing: Quantitative Corpus Linguistic Approaches*. Cambridge University Press.

Finlayson, M.; Mueller, A.; Gehrmann, S.; Shieber, S.; Linzen, T.; and Belinkov, Y. 2021. Causal Analysis of Syntactic Agreement Mechanisms in Neural Language Models. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1828–1843. Online: Association for Computational Linguistics.

Futrell, R.; and Mahowald, K. 2025. How Linguistics Learned to Stop Worrying and Love the Language Models. *arXiv preprint arXiv:2501.17047*.

Hao, Y.; Jin, Z.; Yang, Q.; Wang, X.; and Liu, H. 2023. To Predict L2 Writing Quality Using Lexical Richness Indices: An Investigation of Learners of Chinese as A Foreign Language. *System*, 118: 103123.

Hao, Y.; Wang, X.; Bin, S.; Yang, Q.; and Liu, H. 2024. How Syntactic Complexity Indices Predict Chinese L2 Writing Quality: An Analysis of Unified Dependency Syntactically-annotated Corpus. *Assessing Writing*, 61: 100847.

Hao, Y.; Wang, X.; and Lin, Y. 2022. Dependency Distance and Its Probability Distribution: Are They the Universals for Measuring Second Language Learners' Language Proficiency? *Journal of Quantitative Linguistics*, 29(4): 485–509.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Hu, M. Y.; Mueller, A.; Ross, C.; Williams, A.; Linzen, T.; Zhuang, C.; Cotterell, R.; Choshen, L.; Warstadt, A.; and Wilcox, E. G. 2024. Findings of the Second BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. In Hu, M. Y.; Mueller, A.; Ross, C.; Williams, A.; Linzen, T.; Zhuang, C.; Choshen, L.; Cotterell, R.; Warstadt, A.; and Wilcox, E. G., eds., *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, 1–21. Miami, FL, USA: Association for Computational Linguistics.

Huang, C. 2015. On Syntactic Analyticity and Parametric Theory.

Huang, Y.; Bai, Y.; Zhu, Z.; Zhang, J.; Zhang, J.; Su, T.; Liu, J.; Lv, C.; Zhang, Y.; Fu, Y.; et al. 2023. C-eval: A Multi-level Multi-discipline Chinese Evaluation Suite for Foundation Models. *Advances in Neural Information Processing Systems*, 36: 62991–63010.

Krashen, S. 1982. Principles and Practice in Second Language Acquisition.

Kyle, K.; Sung, H.; Eguchi, M.; and Zenker, F. 2024. Evaluating Evidence for the Reliability and Validity of Lexical Diversity Indices in L2 Oral Task Responses. *Studies in Second Language Acquisition*, 46(1): 278–299.

Lenneberg, E. H. 1967. The Biological Foundations of Language. *Hospital Practice*, 2(12): 59–67.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024a. Deepseek-v3 Technical Report. *arXiv preprint arXiv:2412.19437*.

Liu, H. 2008. Dependency Distance as A Metric of Language Comprehension Difficulty. *Journal of Cognitive Science*, 9(2): 159–191.

Liu, Y.; Shen, Y.; Zhu, H.; Xu, L.; Qian, Z.; Song, S.; Zhang, K.; Tang, J.; Zhang, P.; Yang, B.; et al. 2024b. Zhoblimp: A Systematic Assessment of Language Models with Linguistic Minimal Pairs in Chinese. *arXiv preprint arXiv:2411.06096*.

McCoy, R. T.; Frank, R.; and Linzen, T. 2020. Does Syntax Need to Grow on Trees? Sources of Hierarchical Inductive Bias in Sequence-to-sequence Networks. *Transactions of the Association for Computational Linguistics*, 8: 125–140.

McCoy, R. T.; and Griffiths, T. L. 2025. Modeling Rapid Language Learning by Distilling Bayesian Priors into Artificial Neural Networks. *Nature communications*, 16(1): 4676.

Oba, M.; Kuribayashi, T.; Ouchi, H.; and Watanabe, T. 2023. Second Language Acquisition of Neural Language Models. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 13557–13572. Toronto, Canada: Association for Computational Linguistics.

Peng, Y.; Yan, W.; and Cheng, L. 2021. Hanyu Shuiping Kaoshi (HSK): A Multi-level, Multi-purpose Proficiency Test. *Language Testing*, 38(2): 326–337.

Pinker, S.; and Prince, A. 1988. On Language and Connectionism: Analysis of A Parallel Distributed Processing Model of Language Acquisition. *Cognition*, 28(1-2): 73–193.

Ramesh, D.; and Sanampudi, S. K. 2022. An Automated Essay Scoring Systems: A Systematic Literature Review. *Artificial Intelligence Review*, 55(3): 2495–2527.

Ravfogel, S.; Elazar, Y.; Gonen, H.; Twiton, M.; and Goldberg, Y. 2020. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7237–7256. Online: Association for Computational Linguistics.

Rumelhart, D. E.; and McClelland, J. L. 1985. On Learning the Past Tenses of English Verbs. Technical report.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*.

Wang, Y.; and Hu, R. 2021. A Prompt-independent and Interpretable Automated Essay Scoring Method for Chinese Second Language Writing. In *China National Conference on Chinese Computational Linguistics*, 450–470. Springer.

Warstadt, A.; and Bowman, S. R. 2020. Can Neural Networks Acquire A Structural Bias From Raw Linguistic Data? In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Warstadt, A.; and Bowman, S. R. 2022. What Artificial Neural Networks Can Tell Us about Human Language Acquisition. *Algebraic Structures in Natural Language*, 17.

Warstadt, A.; Mueller, A.; Choshen, L.; Wilcox, E.; Zhuang, C.; Ciro, J.; Mosquera, R.; Paranjabe, B.; Williams, A.; Linzen, T.; and Cotterell, R. 2023. Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. In Warstadt, A.; Mueller, A.; Choshen, L.; Wilcox, E.; Zhuang, C.; Ciro, J.; Mosquera, R.; Paranjabe, B.; Williams, A.; Linzen, T.; and Cotterell, R., eds., *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, 1–34. Singapore: Association for Computational Linguistics.

Warstadt, A.; Zhang, Y.; Li, X.; Liu, H.; and Bowman, S. R. 2020. Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations (Eventually). In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 217–235. Online: Association for Computational Linguistics.

Yan, X.; and Lin, J. 2023. Chinese Character Matters!: An Examination of Linguistic Accuracy in Writing Performances on the HSK Test. *Assessing Writing*, 57: 100767.

Yedetore, A.; Linzen, T.; Frank, R.; and McCoy, R. T. 2023. How Poor is the Stimulus? Evaluating Hierarchical Generalization in Neural Networks Trained on Child-directed Speech. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9370–9393. Toronto, Canada: Association for Computational Linguistics.

Zheng, Y.; Zhang, R.; Zhang, J.; Ye, Y.; and Luo, Z. 2024. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. In Cao, Y.; Feng, Y.; and Xiong, D., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 400–410. Bangkok, Thailand: Association for Computational Linguistics.