Nine-year-old children outperformed ChatGPT in emotion: Evidence from Chinese writing

Siyi Cao[12], Tongquan Zhou[1*], Siruo Zhou[3*]

1. School of Foreign Languages, Southeast University, Nanjing, China, 211189
2. Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, China
3. School of Foreign Studies, Nanjing University of Posts and Telecommunications, Nanjing, China 210023

* Corresponding authors: zhoutongquan@126.com, susanzhou@naver.com

Abstract

ChatGPT has been demonstrated to possess significant capabilities in generating intricate, human-like text, and recent studies have established that its performance in theory of mind tasks is comparable to that of a nine-year-old child. However, it remains uncertain whether ChatGPT surpasses nine-year-old children in Chinese writing proficiency. To explore this, our study juxtaposed the Chinese writing performance of ChatGPT and nine-year-old children on both narrative and scientific topics, aiming to uncover the relative strengths and weaknesses of ChatGPT in writing.

The collected data were analyzed across five linguistic dimensions: fluency, accuracy, complexity, cohesion, and emotion. Each dimension underwent assessment through precise indices. The findings revealed that nine-year-old children excelled beyond ChatGPT in terms of fluency and cohesion within their writing. In contrast, ChatGPT manifested a superior performance in accuracy compared to the children. Concerning complexity, children exhibited superior skills in science-themed writing, while ChatGPT prevailed in nature-themed writing. Significantly, this research is pioneering in revealing that nine-year-old children convey stronger emotions than ChatGPT in their Chinese compositions.

1. Introduction

Artificial intelligence (AI) has shown impressive growth and diversity in recent years, with particular strides being made in the field of natural language learning (Dergaa et al., 2023). Among the vanguards in this arena, OpenAI's ChatGPT stands out with its compelling capabilities to generate intricate, human-like text (De Angelis et al., 2023). This has ignited a fascinating debate about the scope and boundaries of AI's language proficiency (Zhou et al., 2023). Notably, research has revealed that Chinese writing is not ChatGPT's strong suit, but its performance on theory of mind (ToM) tasks is comparable to that of a nine-year-old child (Kosinski, 2023). Within this fascinating context, the current research seeks to probe into an intriguing yet

underexplored area: how well can AI, exemplified by ChatGPT, perform in writing Chinese essays compared to nine-year-old native Chinese speakers?

This study sets out to fill this research void by drawing a comparison between the writing quality of ChatGPT and nine-year-old Chinese native speakers. We will consider five critical dimensions of language performance: fluency, accuracy, complexity, cohesion, and emotion. Additionally, each of these dimensions will be individually assessed by essential indices. Our objective is to shed light on the areas where ChatGPT excels or falls short in comparison to human writers, particularly nine-year-olds, in the context of Chinese writing.

1.1 ChatGPT's development and language capability

ChatGPT is premised on the transformative GPT-4 model, which is a deep learning architecture designed for understanding and generating natural language. The model's training involves two core steps: a generative, unsupervised pre-training phase, which exposes the model to large quantities of unlabeled data, followed by a discriminative, supervised fine-tuning stage that optimizes the model for specific tasks (Chung et al., 2021). This training approach has equipped ChatGPT with the proficiency to perform a variety of tasks such as automated summarization (Ray, 2023), machine translation (Peng et al., 2023), and question-answering (Tan et al., 2023).

ChatGPT demonstrates impressive conversational abilities due to its training process and architecture (Transformer & Zhavoronkov, 2022). Trained on a vast dataset of online texts spanning diverse topics and genres, ChatGPT can comprehend queries and generate coherent, human-like responses (Ray, 2023). Furthermore, reinforcement learning techniques enabled the model to refine its skills through human feedback (Liu et al., 2023). This allows ChatGPT to understand user intent, adjust its responses accordingly, and sustain context-aware dialogues. Through this specialized training regime, ChatGPT has become adept at natural conversation, showcasing advanced language understanding and generation capabilities (Vaishya et al., 2023).

While ChatGPT has garnered attention for its impressive capabilities and practical applications, it has also stirred debates in the linguistics sphere. Distinguished linguists such as Noam Chomsky expressed concerns that ChatGPT's approach devalues humans' innate *linguistic competence*. According to Chomsky, humans intuitively understand and form grammatically accurate sentences (Collins, 2007), suggesting that our language competence is a fundamental cognitive function, not mere imitation (McGilvray, 2014). Chomsky's criticism stems from ChatGPT's role as a data duplicator, a stance at odds with his viewpoint (Chomsky et al., 2023). Similarly, Steven Pinker advocated for a '*language instinct*,' asserting that our innate, genetically coded language learning ability distinguishes humans from other species (Pinker, 2003). Yulin Yuan added to this discourse, arguing that ChatGPT was ill-suited as a technological interface for human-robot interaction. He believed that effective language usage demands embodied intelligence, incorporating lexical

grounding and environmental affordances in embodied cognition (Yuan, 2023). According to these theoretical views, ChatGPT, while able to replicate data patterns, lacks the genuine linguistic intelligence required for language acquisition and use.

Contrasting prior critiques, recent studies highlighted ChatGPT's exceptional language capabilities. Jiao et al. (2023) found that ChatGPT's translation skills compete with established commercial products like Google, DeepL, and Tencent, affirming its prowess as a translator. In terms of discourse skills, Benzon (2023) identified that ChatGPT could engage in analogical reasoning, interpret films and stories, and comprehend abstract concepts like justice and charity. Additionally, it could modulate its discourse level to match children of different ages. On language usage, Cai et al. (2023) assessed 12 tasks involving sound, words, syntax, meaning, discourse, and interlocutor sensitivity. They discovered that ChatGPT's choices paralleled human decisions in 10 tasks, though it did not favor shorter words for less informative content and lacked context utilization to clear syntactic ambiguities. Regarding psychological-linguistic skills, Kosinski (2023) found that ChatGPT's performance on theory of mind (ToM) tasks was akin to a nine-year-old child. For instance, given a narrative involving the relocation of a cat, ChatGPT accurately deduced the cat's location. ChatGPT notably achieved 100% accuracy across all 20 tasks, showcasing advanced cognitive abilities. These findings pose the intriguing question of whether ChatGPT's writing capabilities, requiring high cognitive abilities, might match those of nine-year-old children.

1.2 Chinese writing by ChatGPT and nine-year-old children

Writing, deemed a pivotal skill in our everyday lives (Nasser, 2016), has undergone a significant transformation with the introduction of AI authors like ChatGPT last year. Previously, the writing domain was solely occupied by human writers; this landscape, however, dramatically shifted with the advent of AI. As highlighted in numerous academic works, ChatGPT promises to deliver an array of benefits such as language support, translation, editing, and proofreading (Shahriar & Hayawi, 2023). In practice, ChatGPT can augment researchers' linguistic proficiency and writing abilities by providing immediate feedback on grammar, syntax, spelling, and vocabulary, consequently enhancing the overall standard of their manuscripts (Zhou et al., 2023). Additionally, ChatGPT can rapidly produce text on intricate or technical subjects. However, it's important to remember Chomsky's observation that this could potentially be characterized as high-tech plagiarism (Chomsky et al., 2023).

The proficiency of ChatGPT in Chinese writing has been conjectured to fall substantially short of its capabilities in high-resource European languages such as English and German, given the status of Chinese as a low-resource or distant language (Jiao et al., 2023). A recent study, seemingly validating this assumption, challenged ChatGPT to craft a fictional Chinese composition based on the prompt, "*请依照题目完成以下的作文'你曾经犯下一个错误，伤透了父母的心。写出犯错经过和你的后悔之情。(You once made a mistake that deeply hurt your parents. Write about the process, your feelings of regret, and the lessons learned)*" (Rudolph et al., 2023).

Although ChatGPT created a response consistent with the topic, the essay was found to lack structure and was riddled with grammatical errors, suggesting a deficiency in its handling of the Chinese language. Nevertheless, the study's limitations include a singular sample and a lack of statistical analysis, the true proficiency of ChatGPT in Chinese writing therefore remains an open question.

Chinese writing is generally considered an important and difficult skill for nine-year-old children (third grader in China), as they have just embarked on learning how to write in this intricate language and possess limited life experiences (Wang, 2021). At this developmental stage, their writing often features simple narrative constructs, leading to compositions that may lack depth and complexity (Zhang, 2023). It's noted that these young students frequently make a multitude of errors in their compositions, including lexical and syntactic mistakes (Sun, 2023). Moreover, the increased occurrence of homophonic and similar-shaped characters contributes to typographical errors, making accurate recognition of Chinese characters more challenging (Wang, 2023). They also commonly make mechanical errors, such as misplacing periods or commas (Yan et al., 2012). Interestingly, as mentioned in 1.1, the advanced cognitive capabilities of ChatGPT, including the theory of mind (ToM), parallel those of a nine-year-old child. This fact raises an intriguing question: could ChatGPT outperform these young Chinese students in terms of Chinese writing, a task requiring advanced cognitive abilities?

Narrative writing, characterized by its vivid depiction of events, experiences, or emotions, demands a creative approach towards elements like characters, plot, setting, and dialogue. This method of writing provides an avenue for individual expression and entertainment (Oatley, 1995). The art of narrative writing enables authors to demonstrate their cognitive capabilities as they delve into character motivations and perspectives. It also necessitates thoughtful consideration and logical reasoning, with special attention to aspects such as plot progression and character development (Prado et al., 2015). Further, a high level of language proficiency and an aptitude for descriptive, captivating language are essential components of narrative writing (McFadden & Gillam, 1996). When compared to other intricate writing tasks, such as argumentative writing, narrative writing seems to be a more accessible and engaging way for nine-year-olds to display their critical thinking and language abilities (Xu, 2018). Thus, it provides a sensible benchmark to evaluate and gauge the writing competency of both nine-year-old children and ChatGPT.

1.3 Assessment of Chinese writing

Thus far, only a limited number of studies have established a scoring method for assessing the writing compositions of young Chinese children. Building upon the English scoring foundation set by Wagner et al. (2011), Yan et al. (2012) devised a scoring framework specifically for 9-year-old Chinese children in Hong Kong. This framework is segmented into two main dimensions: content and organization. The content dimension evaluates relevance, breadth, and depth, while organization assesses sentence and paragraph structures, key elements, and intelligibility. Drawing

inspiration from Yan et al. (2012), Tong et al. (2014) introduced a modified scoring model for 11-year-olds in Beijing, incorporating criteria for grammatical errors and punctuation accuracy. Yet, current methods neglect vital aspects of writing, such as emotional intensity. Given these gaps, our present study proposes a novel scoring approach tailored to the writing needs of young Chinese children.

The current study introduces a scoring method that spans five dimensions: fluency, accuracy, complexity, cohesion, and emotion (Table 1). Within the fluency dimension, evaluation is based on total number of sentences and T-units. Cahyono et al. (2016) have found a strong correlation between the total number of sentences and writing fluency, underscoring its validity as a measure. Additionally, A T-unit is characterized as an independent clause accompanied by any subordinate elements linked to it, whether they be phrases or clauses (Wagner et al., 2011). Yan et al. (2012) raised concerns about the applicability of the T-unit in assessing Chinese writing due to the fluid and occasionally ambiguous nature of Chinese punctuation, particularly in comparison to English punctuation norms. However, Jiang (2013) validated the T-unit as a trustworthy tool for scrutinizing the progression of Chinese writing.

Within the accuracy dimension, evaluation relies on the count of spelling errors, grammatical errors, and punctuation errors, as posited by Tong et al. (2014). For clarity, grammatical errors encompass issues like disordered word sequences, fragmented sentences, misnomers, collocation mistakes, redundancy, and ambiguous or incorrect usage. Regarding the complexity dimension, three metrics were employed to assess the quality of Chinese writing: the frequency of uncommon words, the number of idioms and the count of unrepeated words. Nippold (2000) contended that the prevalence of uncommon words and idioms directly correlates with an article's complexity; a higher frequency tends to render articles more challenging to comprehend . Moreover, Isaacson (1988) suggested that the number of unrepeated words in a text reflects vocabulary diversity. A greater number of unrepeated words points to higher lexical complexity, signaling a broader vocabulary range.

In the cohesion dimension, two factors are considered: the total number of connectives and the number of accurately used connectives. Crossley & McNamara (2012) suggested that the use of connectives is integral to the cohesion of a written piece. Regarding the emotion dimension, the evaluation just focused only one main factor: the intensity of emotion. As pointed out by Carlbring et al. (2023), machines currently lack the depth of human empathy and emotion. As a result, their writings might not exhibit strong emotional intensity. This could be a distinguishing factor between human and AI-generated writing.

Table 1. The description of five dimensions

Dimensions	**Indicators**	**Description**
Fluency	total number of sentences	TNOS
	T-units	TUNIT
Accuracy	the number of spelling errors	NOSE
	the number of grammatical errors	NOGE
	the number of punctuation errors	NOPE

Complexity	the number of uncommon words	NOUCW
	the number of idioms	NUI
	the number of unrepeated words	NOURW
Cohesion	the number of correct connectives	NOCC
	total number of connectives	TNOC
Emotion	the intensity of emotion	IOE

In summary, there are unresolved questions regarding ChatGPT's capabilities. First, while theoretical studies have highlighted a potential weakness in ChatGPT's handling of the Chinese language, a statistical analysis has yet to determine its true proficiency in Chinese writing. Second, even though ChatGPT's performance on ToM tasks is on par with that of a nine-year-old child, it remains uncertain whether ChatGPT surpasses nine-year-olds in Chinese narrative writing. This study seeks to address the following research questions concerning ChatGPT and nine-year-old children in the context of Chinese writing:

(a) Does ChatGPT outperform nine-year-old children in narrative writing?

(b) How do the performances of ChatGPT and nine-year-old children compare across five dimensions: fluency, accuracy, complexity, cohesion, and emotion?

2. Method

2.1 Participants

In our study, 30 nine-year-old children (third-grader) from a primary school voluntarily participated, comprising 14 females and 16 males. These children, aged between 8 and 10 (M = 9.04, SD = 0.73). The experiment received approval from the Human Research Ethics Committee of the university affiliated with the first author.

2.2 Writing tasks

All 30 nine-year-old participants were instructed to compose two narrative essays on specific themes (i.e., natural and scientific) related to Chinese writing, resulting in 60 articles in total.

The first topic was centered on nature: "*这一单元，我们看到了作者笔下一幅幅动人的面面，那乡下的动物、植物、人家，那孩子眼中小小的神奇的天窗……这些画面中都饱含着作者的情感。请以《我心上的这幅画》为题,写一篇习作,表达自己的真情实感。300 字左右。(In this unit, we witnessed a series of touching scenes penned by the author: the animals, plants, and households of the countryside, and the magical little skylight in a child's eyes... Each of these images is imbued with the author's emotions. Please write a composition titled "The Painting in My Heart", expressing your genuine feelings. It should be around 300 words.*)".

The second topic revolved around science: "*科学离不开生活，科技的进步来源于人类对美好生活的想象创造。未来的你如果是一名科技工作者，你想发明什么?它是什么样子的?有哪些功能?让我们把它写出来吧!请完成习作，300 字左右*

(Science is inseparable from life, and the advancement of technology stems from humanity's imagination and creation of a better life. If you become a technologist in the future, what would you like to invent? What does it look like? What functions does it have? Let's write it down! Please complete the assignment, approximately 300 words.)".

Participants were allocated 80 minutes (the duration of two classes), to complete their submissions. For statistical evaluation, the first author digitized all handwritten essays. To guarantee consistency between the handwritten and digital versions, another laboratory member verified the congruence of both forms. Additionally, for comparative analysis, ChatGPT was employed to produce 60 unique essays on these themes with the same prompt, ensuring no repetition in content.

2.3 Data collection and coding

A total of 120 Chinese texts (30*2*2) were collected from nine-year-old children and ChatGPT under natural and scientific topics. To construct an assessment model encompassing five linguistic dimensions, data for each index within every dimension was meticulously gathered. On one hand, indices including total number of sentences (TNOS), number of spelling errors (NOSE), number of punctuation errors (NOPE), number of uncommon words (NOUCW), number of idioms (NUI), and intensity of Emotion (IOE) were extracted from https://xiezuocat.com/. This website serves as a valuable resource for writing assessment, offering scores across various dimensions to enhance writing quality. Additionally, it furnishes emotion assessments in Chinese, primarily leveraging natural language processing techniques. To validate these scores, two lab members cross-verified the outcomes supplied by the website.

On the other hand, indices like T-units (TUNIT), number of grammatical Errors (NOGE), number of unrepeated words (NOURW), number of correct connectives (NOCC), and total number of connectives (TNOC) were collected by an additional two lab members. Inter-rater reliabilities for these elements were calculated based on 120 samples and manifested as Pearson's correlation coefficients, ranging between .76 and .89. The final score was determined by averaging the scores from both members.

2.4 Data analysis

Firstly, we conducted a Confirmatory Factor Analysis (CFA) to validate a predefined assessment model for Chinese writing across five dimensions, using the "lavaan" package in R (R Core Team, 2016). This was followed by testing the data's fit to the model to affirm its validity. Specifically, the CFA sought to determine if the 11 linguistic factors could be effectively predicted by five measured variables. The goodness-of-fit for the specified model was gauged using multiple statistical tests and indices. The chi-square test, for instance, was employed to assess the null hypothesis that the covariance matrix from our model aligns with the observed data covariance matrix. Ideally, for this test, we aimed for a non-significant result ($p > .05$) to uphold

the null hypothesis. Given our relatively small sample size ($N = 120$), we favored the CFI and SRMR to judge the model's fit, as the RMSEA and TLI can be overly critical in smaller samples. With our predetermined criteria, we considered models exhibiting a CFI ≥ .90 and SRMR ≤ .07 to have a satisfactory fit.

Secondly, a structural equation model (SEM), also accessed by the "lavaan" package in R, was applied to build a model to see the relationship between "group" (nine-year-old children vs. ChatGPT), "writing type" (science vs. nature) and five latent measured factors: fluency, accuracy, complexity, cohesion and emotion. Given the limited sample size ($N = 120$), we relied on the CFI and SRMR to evaluate model fit. This is because both RMSEA and TLI tend to inaccurately reject models when the sample size is small. Based on our predetermined criteria, a model was considered to have an adequate fit if the CFI was ≥ .90 and the SRMR was ≤ .07.

Thirdly, we conducted two-way analyses of variances (ANOVAs) to discern differences across the five linguistic dimensions based on "group" and "writing type". For this, we used the EMMEANS function from the bruceR package (Bao, 2023). Whenever a significant main effect emerged, we performed multiple comparisons using the Tukey method.

3. Results

The CFA modeled five latent linguistic dimensions for a total of 11 factors, which means that each linguistic dimension (i.e., fluency, accuracy, complexity, cohesion and emotion) could be predicted by various separate linguistic variables. The results showed that the model had adequate fit based on a priori criteria ($\chi^2 = 91.6$, $df = 35$, CFI = .93, SRMR = .07) and factor loadings were detailed in Table 2.

Table 2. Factor loadings for model indices

	λ	***S.E.***	***p***
Fluency			
TNOS	.49	.50	< .001
TUNIT	.93	.82	< .001
Accuracy			
NOSE	.93	.34	< .001
NOGE	.56	.07	< .001
NOPE	.39	.22	< .001
Complexity			
NOUCW	.93	1.7	< .001
NUI	.76	1.29	< .001
NOURW	.27	.08	< .001
Cohesion			
NOCC	1.0	.27	< .001
TNOC	.99	.26	< .001
Emotion			

IOE	1.0	.01	< .001

Additionally, the SEM modeled the variables of "group" (nine-year-old children vs. ChatGPT) and "writing type" (science vs. nature) as predictors of five latent linguistic dimensions. The model is presented in Fig 1. The model had adequate fit based on set a priori criteria (χ^2 = 146.71, *df* = 47, CFI = .91, SRMR = .07). It means that the variables of "group" and "writing type" were significantly associated with five latent linguistic dimensions. Specifically, "group" was significant positively associated with fluency (B = .25, p < .05), accuracy (B = .92, p < 0.001) and emotion (B = .39, p < 0.001). In contrast, "writing type" was significant negatively associated with complexity (B = -.55, p < 0.001), cohesion (B = -.64, p < 0.001) and emotion (B = -.49, p < 0.001).

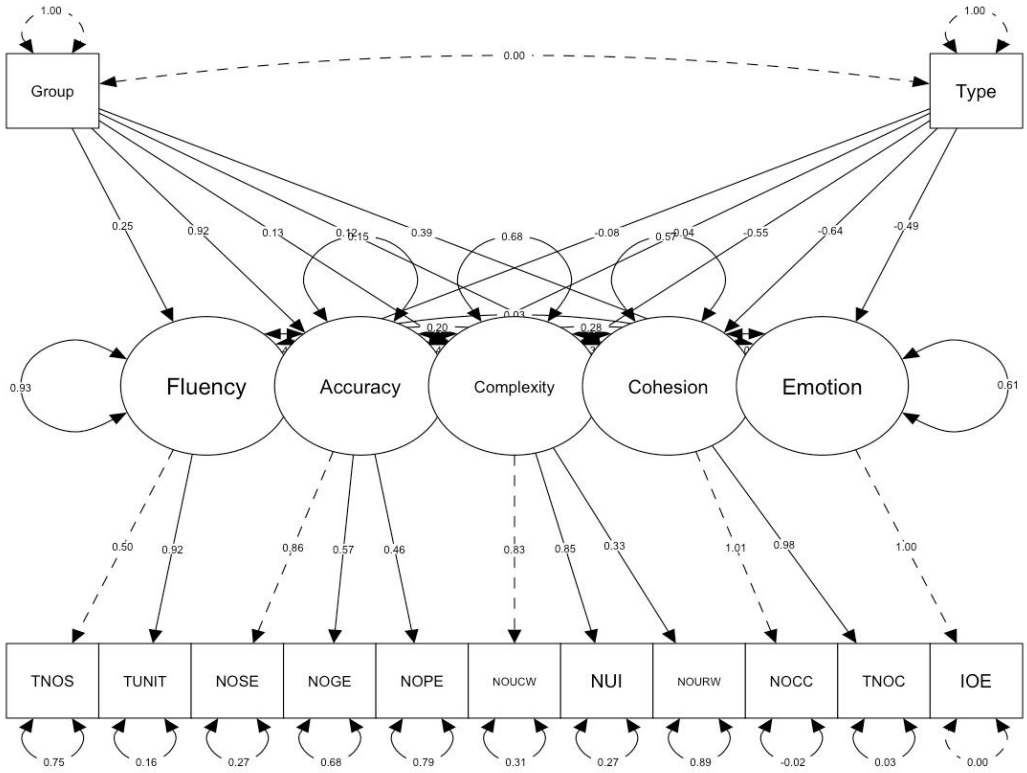

Figure 1. Structural equation model with "Group" and "Type" as predictors of five linguistic factors. *Note* All parameters estimates are standardized. Dashed lines represents the indicators that are fixed in the fixed loading method.

Two-way analyses of variance (ANOVAs) were conducted to assess whether ChatGPT outperformed nine-year-old children on five linguistic factors across different types of Chinese writing (Table 3). The factors examined included fluency, accuracy, complexity, cohesion, and emotion, which served as the dependent variables. The independent variables were "Group" (i.e., nine-year-old children vs. ChatGPT) and "Writing Type" (i.e., science vs. nature).

For fluency, a significant main effect was observed for the "Group" variable ($F(1, 116) = 8.623$, $p < .01$), but no significant interaction between "Group" and "Writing Type" was found ($F(1, 116) = .78$, $p = .379$). Pairwise comparisons revealed that nine-year-old children exhibited higher fluency levels than ChatGPT in science-themed writing (β(nine-year-old children–ChatGPT) = 1.689, $t(116) = 2.70$, $p < .01$).

Regarding accuracy, both a significant main effect for "Group" ($F(1, 116) = 226.15$, $p < .001$) and a significant interaction between "Group" and "Writing Type" were noted ($F(1, 116) = 8.30$, $p < .01$). Pairwise comparisons indicated that in nature-themed writing, nine-year-old children demonstrated higher values than ChatGPT (β(nine-year-old children–ChatGPT) = 3.514, $t(116) = 8.60$, $p < .001$). Similarly, in science-themed writing, nine-year-old children also outperformed ChatGPT in terms of values (β(nine-year-old children–ChatGPT) = 5.17, $t(116) = 12.67$, $p < .001$). These accuracy values were derived from the counts of three different types of errors. In this context, higher values imply lower accuracy. Consequently, the results suggest that ChatGPT exhibited greater accuracy than nine-year-old children across both types of writing.

In terms of complexity, two significant main effects were observed: one for the "Group" variable ($F(1, 116) = 4.62$, $p < .05$) and another for "Writing Type" ($F(1, 116) = 18.59$, $p < .001$). Additionally, a significant interaction between "Group" and "Writing Type" was noted ($F(1, 116) = 5.79$, $p < .05$). Pairwise comparisons revealed that in the domain of science-themed writing, nine-year-old children displayed greater complexity than ChatGPT (β(nine-year-old children–ChatGPT) = 14.78, $t(116) = 3.22$, $p < .01$). Intriguingly, ChatGPT exhibited higher complexity in nature-themed writing compared to its performance in science-themed writing (β(science–nature) = -21.80, $t(116) = -4.750$, $p < .001$).

In the aspect of cohesion, there were two significant main effects of "Group" ($F(1, 116) = 4.10$, $p < .05$) and "Writing Type" ($F(1, 116) = 92.53$, $p < .001$). Furthermore, a significant interaction effect between "Group" and "Writing Type" was also observed ($F(1, 116) = 9.17$, $p < .001$). Pairwise comparisons highlighted that in science-themed writing, nine-year-old children demonstrated greater cohesion than ChatGPT (β(nine-year-old children–ChatGPT) = 2.864, $t(116) = 3.57$, $p < .001$). Interestingly, both ChatGPT and the nine-year-old children exhibited higher levels of cohesion in nature-themed writing as compared to their performance in science-themed writing (ChatGPT: β(science–nature) = -7.17, $t(116) = -8.94$, $p < .001$; nine-year-old children: β(science–nature) = -3.74, $t(116) = -4.66$, $p < .001$).

Regarding the emotion factor, two significant main effects were identified: one for the "Group" variable ($F(1, 116) = 34.82$, $p < .001$) and another for the "Writing Type" ($F(1, 116) = 54.62$, $p < .001$). Additionally, a notable interaction effect between "Group" and "Writing Type" was observed ($F(1, 116) = 23.26$, $p < .001$). Pairwise comparisons illuminated that nine-year-old children displayed greater emotional depth than ChatGPT in both science-themed and nature-themed writing (β(nine-year-old children–ChatGPT) = .182, $t(116) = 7.58$, $p < .001$; β(nine-year-old children–ChatGPT) = .22, $t(116) = 9.40$, $p < .001$). Intriguingly, ChatGPT manifested

higher levels of emotion in nature-themed writing compared to its performance in science-themed writing (β(science–nature) = -2.21, t(116) = -8.63, p < .001). However, for the nine-year-old children, there was no significant difference in emotional expression between nature-themed and science-themed writing (β(science–nature) = -.04, t(116) = -1.82, p = .07).

4. Discussion

The current study compared the quality of Chinese writing generated by ChatGPT and nine-year-old children in terms of five critical dimensions of language performance: fluency, accuracy, complexity, cohesion, and emotion. The results revealed that nine-year-old children outperformed ChatGPT in terms of fluency, cohesion, and emotion in their writing. Conversely, ChatGPT demonstrated superior performance over nine-year-old children in terms of accuracy. Regarding complexity, nine-year-old children showed better performance than ChatGPT in science-themed writing, but this result reversed in nature-themed writing.

In the following sections, we will delve into potential factors that may have contributed to these results across the five discourse components and their correlations in narrative writing.

4.1 Fluency: nine-year-old children > ChatGPT

The statistical analysis revealed that nine-year-old children exhibited greater fluency than ChatGPT, as measured by the total number of sentences (TNOS) and T-units (TUNIT). According to the further analysis, nine-year-old children have more TNOS (p < .05) and TUNIT (p < .01). This outcome can be attributed to two potential reasons. Firstly, the length of Chinese compositions produced by ChatGPT frequently fell short of the 300-word benchmark, suggesting a possible misunderstanding of the prompt "*300 字左右*" (meaning 'around 300 words'). While "*左右*" translates to "around" in English, ChatGPT might not interpret it accurately within the context of the prompt. To validate this hypothesis, the researchers presented several analogous prompts like "*200 字左右*" (meaning 'around 200 words') and "*500 字左右*" (meaning 'around 500 words') to ChatGPT. The resulting text was consistently either significantly shorter than, or exceeded, the 200/500-word guidelines. This observation suggests a potential oversight by ChatGPT's developers regarding the term "*左右*" (meaning 'around') in Chinese. Further research is warranted to identify and address this specific challenge.

Secondly, data-training limitations of ChatGPT leads to fewer T-units than nine-year-old children. In fact, ChatGPT's responses are shaped by its vast training data, restricting its ability to generate innovative ideas (Abdullah & Jararweh, 2022). As a result, ChatGPT often generates sentences that are directly related to the topic at hand, such as "*这幅画中有着美丽的乡村风景和活泼可爱的农村孩子 (This painting features beautiful countryside scenery and lively, adorable rural children)*", without venturing into deeper or more imaginative descriptions. In contrast,

nine-year-old children frequently craft intriguing sentences, reflecting their imaginative capabilities and personal experiences. For instance, in a nature-themed essay, a 9-year-old described not just the act of climbing but also the breathtaking view upon reaching the top, writing, "*我马不停蹄的往观景台那去，看到了比火烧云还美的一景 (I hastily headed towards the observation deck and saw a scene even more beautiful than the burning clouds)*". From this, it's evident that children's writing contains a multitude of sub-events under a central theme, whereas ChatGPT's output might be more narrowly focused. Given that a T-unit consists of an independent clause along with any related subordinate elements, the number of T-units is likely associated with the number of events described. As such, nine-year-old children tend to produce more T-units than ChatGPT.

4.2 Accuracy: ChatGPT > nine-year-old children

The study's findings indicated that ChatGPT outperformed nine-year-old children in terms of accuracy, a metric assessed by the number of spelling errors (NOSE), grammatical errors (NOGE), and punctuation errors (NOPE). Subsequent analysis elucidates that ChatGPT commits fewer NOSE ($p < .01$), NOGE ($p < .01$), and also NOPE ($p < .01$) compared to nine-year-old children. Several plausible explanations can be posited to clarify these outcomes.

First, the intrinsic complexity of Chinese orthography significantly contributes to the abundance of spelling errors observed in the writings of nine-year-old Chinese students. Chinese characters operate as logograms; each one symbolizes a word or a meaningful component of a word. In contrast to alphabetic systems, where letters signify sounds, every Chinese character possesses a distinct structure and form. This inherent complexity is a fertile ground for errors as children navigate through the learning process (Sun, 2023). Furthermore, the prevalent existence of homophonic characters and those with similar shapes can induce confusion and result in mistakes for children at this developmental stage (Wang, 2023).

Second, nine-year-old children still struggle with abstract concepts, such as grammatical rules. According to the cognitive developmental theory of Piaget (1976), children undergo a series of developmental stages, with each stage marked by unique cognitive capabilities. At around 9 years of age, children generally find themselves in the "Concrete Operational Stage" of development, a phase extending roughly from 7 to 11 years. In this stage, children start developing a more organized and rational thought process. They can understand basic grammatical structures and rules as they can think logically about concrete events and situations they encounter. However, their understanding is often tied to tangible and visible situations, concepts, and objects, making abstract grammatical rules harder to grasp (Babakr et al., 2019).

Third, nine-year-old children are still in the phase of grasping the application of punctuation, such as commas, and often struggle with utilizing them correctly. As elucidated by Tang (2018), the learning of punctuation during the primary school stage is segmented into three phases. In the initial phase, targeting 1st and 2nd graders, students are introduced to the usage of common symbols found within texts, such as

commas and periods. The subsequent phase focuses on 3rd and 4th graders, progressing to the comprehension of more intricate punctuation like colons and quotation marks. By the final phase, the emphasis shifts to the accurate application of these punctuation marks. Hence, nine-year-olds are typically still in the learning phase and often demonstrate unfamiliarity with appropriate punctuation in their writing. This is evident, for instance, in sentences like "*每当我们去散步总会看见有人把桌子摆到门口吃 (Whenever we go for a walk, we always see people setting up tables at the doorway to eat)*," where punctuation may be used inaccurately. However, ChatGPT is trained by a big quantity of data and it is impossible to have spelling, grammatical and even punctuation data (Bishop, 2023).

4.3 Complexity: nine-year-old children > ChatGPT (sci-themed writing)

The results illustrated that nine-year-old children exhibited superior complexity compared to ChatGPT in science-themed writing; conversely, this outcome was inverted in nature-themed writing. This suggests that nine-year-olds utilized a greater number of uncommon words, idioms, and unrepeated words than ChatGPT in science-themed compositions, but this scenario was reversed in nature-themed writing. A few factors could elucidate the noted disparities in complexity between nine-year-old children and ChatGPT.

Firstly, the increased thematic familiarity that children possess regarding science-themed subjects could be a contributing factor. Children, at the pivotal age of nine, undergo crucial cognitive development and learning stages (Piaget, 1976), engaging with diverse learning materials, including those on scientific subjects. This exposure likely enhances their utilization of uncommon words, idioms, and unrepeated words in science-themed writing. Recent introductions to new scientific concepts may predispose them to employ a diverse range of words and expressions to articulate scientific phenomena.

Secondly, the disparity may emanate from the constraints of ChatGPT's training data. Being trained on a varied spectrum of internet text, ChatGPT may exhibit more proficiency in generating text on themes more prevalently represented in its training data (Dwivedi et al., 2023). The predominance of nature-themed topics in the training data could account for ChatGPT's superior performance in nature-themed writing, showcasing a more affluent vocabulary and a higher utilization of idioms.

Thirdly, the natural inclination of children towards creativity and imagination is noteworthy (Shidiq, 2023). Children's vivid imaginations and inherent creative tendencies can drive them to employ a richer, more varied vocabulary, especially in subjects they find intriguing or have newly acquired knowledge about. This intrinsic creativity and imagination may find a more tangible expression in their science-themed compositions, particularly if they find the subject matter engaging or stimulating.

4.4 Cohesion: nine-year-old children > ChatGPT

According to the statistical analysis, nine-year-old children exhibit superior cohesion to ChatGPT in science-themed compositions. Notably, both groups display enhanced cohesion in nature-themed prose, surpassing their output in science-related pieces.

Cohesion in writing encompasses the seamless and logical integration of linguistic elements like conjunctions. Observations indicate that ChatGPT infrequently employs causal connectives, such as "*所以*" and "*因此*" (meaning 'therefore'), averaging 5.23 connectives per piece. It, however, manifests a propensity for over-utilizing "*并/并且*" (meaning 'and'), a coordinating conjunction, approximately three times per composition. For instance, ChatGPT interconnected two casual sentences using "*并且*" in: "*房前的院子里有一只鹿在吃草，并且看起来非常宁静 (In the yard in front of the house, a deer is grazing, and it appears very peaceful)*," rendering the connection somewhat unusual. Furthermore, ChatGPT predominantly exhibits restricted utilization of conjunctive adverbs like "*即使*" (meaning 'even though') to forge cohesive links between sentences.

Conversely, nine-year-old children leverage a more diversified array of causal connectives, including "*于是*" (therefore), also averaging 5.23 connectives, proficiently employing them to elucidate logical relations between sentences. However, they occasionally integrate more informal connectives, substituting "*结果*" for "*但*" (meaning 'but') as illustrated in: "*我又偷偷的溜到它们后面偷鸭蛋，结果它们叫了一声 (I joyfully sneaked behind them to steal duck eggs, but they quacked loudly as a result)*."

Our findings corroborate those of Zhou et al. (2023), who highlighted that Chinese intermediate English majors outperformed ChatGPT in deep cohesion. These congruent outcomes likely stem from the inherent limitations of text generation models, as elucidated by Zhao et al. (2022). Present-day models notably struggle with coherence, producing texts lacking logical flow and coherence, complicating comprehension for the reader. Despite strides in Natural Language Processing (NLP), the persistence of incoherence in generated texts poses a substantial obstacle. This predicament arises as these models, fundamentally statistical, lack comprehensive understanding of context and the inherent meaning of texts. Consequently, enhancing coherence in text generation remains a pivotal focus in ongoing NLP research endeavors, including those related to ChatGPT.

4.5 Emotion: nine-year-old children > ChatGPT

Intriguingly, our research has made an initial discovery that the intensity of emotional expression in the writing of nine-year-old children surpasses that of ChatGPT in Chinese prose. A plausible explanation for this disparity is ChatGPT's inherent lack of emotions. It fabricates responses by leveraging patterns discerned during its training phase. Thus, its outputs, although seemingly emotive, are devoid of genuine emotional states and simply mirror human-like reactions, as substantiated by Zhao et al. (2022). In contrast, children possess authentic emotions, rendering their expressions in writing more robust and palpable. This is exemplified vividly in

science-themed compositions. For instance, a nine-year-old child employed a tag question to exude pride in their invention, "*那这种书包甚至没有什么重量，假如你是一位大学生，那你的书包就跟小学生的书包差不多，怎么样，是不是特别神奇?" (This kind of backpack doesn't even have much weight. If you are a college student, then your backpack would be similar to that of an elementary school student. How about that, isn't it quite magical?*). Despite the colloquialism, the sentences brim with childlike wonder and pronounced personal emotion.

Conversely, the prose of ChatGPT resembles emotionless tech broadcasts, "*智能环境监测器是我未来的创新发明，它将通过精确的传感技术和智能算法，帮助人们监测和改善室内环境质量，提高生活质量" (The intelligent environmental monitor is my innovative invention for the future. It will employ precise sensing technology and intelligent algorithms to help people monitor and improve indoor environmental quality, thereby enhancing the quality of life*). This stark contrast underscores ChatGPT's mechanical nature, devoid of human emotional nuances.

However, our findings diverge from the study by Zhao et al. (2023), which attributed promising capacities for emotional response generation in dialogues to ChatGPT. This inconsistency could stem from Zhao et al. (2023) juxtaposing the performance of ChatGPT with other supervised baseline models instead of with actual humans. Additionally, the focal point of Zhao et al. (2023) was dialogue, as opposed to our study's emphasis on Chinese prose. Dialogues are typically concise, consisting of single sentences, while prose requires multi-sentence logical coherence. Thus, synthesizing emotionally resonant prose remains a formidable challenge for models like ChatGPT.

Conclusion

The present research undertook a comparative analysis of Chinese compositions produced by ChatGPT and nine-year-old children, focusing on five pivotal aspects of linguistic proficiency: fluency, accuracy, complexity, cohesion, and emotion. The findings unveiled that, in the domains of fluency, cohesion, and emotion, the children's writing excelled compared to ChatGPT. In contrast, ChatGPT surpassed the children in rendering accurate compositions. Concerning complexity, the children exhibited superior capabilities in science-themed compositions, a trend that was inverted in compositions with nature themes.

This research yields significant insights, possessing theoretical and applied relevance for large language models and the methodology of teaching Chinese. It illustrates that, when pitted against nine-year-old children, ChatGPT is highly proficient in crafting accurate compositions but manifests deficiencies in cohesion and emotional expression, underscoring the imperative for future advancements to augment the capabilities of AI-driven writing instruments, especially in forging sophisticated logical links and conveying emotion.

Moreover, educators ought to motivate children of this age to refine the precision of their writing, focusing on aspects like punctuation and grammatical structures to enhance their comprehensive writing proficiency. The linguistic facets of fluency,

accuracy, and complexity could serve as benchmarks to assess the efficacy of writing instruction in classrooms, thereby informing instructional design.

However, our research presents some constraints. Primarily, it is centered on Chinese compositions by nine-year-old children, who displayed more grammatical and punctuation inaccuracies, without exploring ChatGPT's performance against adult native Chinese speakers, who possess advanced linguistic proficiency. Secondly, the investigation is confined to narrative compositions themed around nature and science, omitting other formats like argumentative writings, prose, and poetry in Chinese. Subsequent investigations should broaden the genres of writing analyzed to present a comprehensive view of the comparative performance between ChatGPT and human writers in Chinese compositions.

Table 3.Five linguistic dimensions of Chinese writing from ChatGPT and 9-year-old children across nature and science-themed writing type

	ChatGPT		9-year-old children					
	Na-themed	Sci-themed	Na-themed	Sci-themed		Group	Writing Type	Group x Writing Type
Fluency	- .21	-1.09	.70	.60	*F-value*	8.62	1.24	.78
					p	< .01	.27	.38
					η^2p	.07	.01	.01
Accuracy	-1.62	-2.73	1.90	2.45	*F-value*	226.15	.92	8.30
					p	< .001	.34	< .01
					η^2p	.66	.01	.07
Complexity	7.41	-14.39	6.58	.39	*F-value*	4.61	18.59	5.79
					p	< .05	< .001	< .05
					η^2p	.04	.14	.05
Cohesion	3.01	-4.16	2.44	-1.30	*F-value*	4.11	92.53	9.17
					p	< .05	< .001	< .01
					η^2p	.03	.44	.07
Emotion	.05	- .15	.07	.03	*F-value*	34.82	54.62	23.26
					p	< .001	< .001	< .001
					η^2p	.23	.32	.17

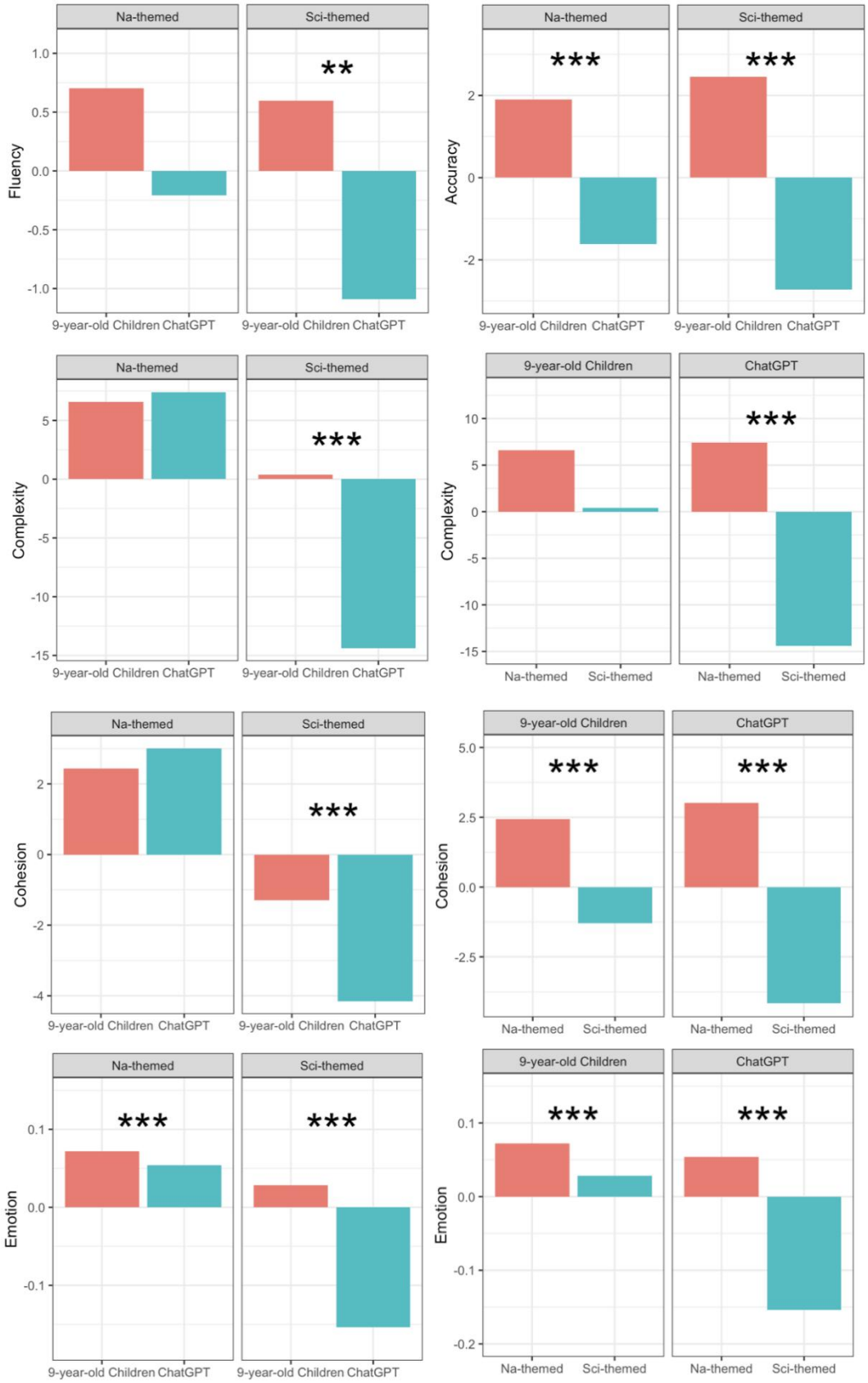

Figure 2. Five linguistic dimensions of Chinese writing produced by ChatGPT and 9-year-old children

Reference

Abdullah, M., Madain, A., & Jararweh, Y. (2022, November). ChatGPT: Fundamentals, applications and social impacts. In *2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 1-8). IEEE.

Babakr, Z., Mohamedamin, P., & Kakamad, K. (2019). Piaget's cognitive developmental theory: Critical review. *Education Quarterly Reviews*, 2(3).

Benzon, W. L. (2023). Discursive Competence in ChatGPT, Part 1: Talking with Dragons.

Bishop, L. (2023). A computer wrote this paper: What chatgpt means for education, research, and writing. *Research, and Writing (January 26, 2023)*.

Cai, Z. G., Haslett, D. A., Duan, X., Wang, S., & Pickering, M. J. (2023). Does ChatGPT resemble humans in language use?. *arXiv preprint arXiv:2303.08014*.

Carlbring, P., Hadjistavropoulos, H., Kleiboer, A., & Andersson, G. (2023). A new era in Internet interventions: The advent of Chat-GPT and AI-assisted therapist guidance. *Internet Interventions*, 32.

Chomsky, N., Roberts, I., & Watumull, J. (2023). Noam Chomsky: The False Promise of ChatGPT. *The New York Times, 8*.

Chung, Y. A., Zhang, Y., Han, W., Chiu, C. C., Qin, J., Pang, R., & Wu, Y. (2021, December). W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. *In 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 244-250). IEEE.

Collins, J. (2007). Linguistic competence without knowledge of language. *Philosophy Compass*, 2(6), 880-895.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2), 115-135.

De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., & Rizzo, C. (2023). ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers in Public Health*, 11, 1166120.

Dergaa, I., Chamari, K., Zmijewski, P., & Saad, H. B. (2023). From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. *Biology of Sport*, 40(2), 615-622.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 102642.

Isaacson, S. (1988). Assessing the writing product: Qualitative and quantitative measures. *Exceptional Children*, 54(6), 528-534.

Jiang, W. (2013). Measurements of development in L2 written production: The case

of L2 Chinese. *Applied Linguistics*, 34(1), 1-24.
Jiao, W., Wang, W., Huang, J. T., Wang, X., & Tu, Z. P. (2023). Is ChatGPT a good translator? Yes with GPT-4 as the engine. *arXiv preprint arXiv:2301.08745*.
Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.
Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., ... & Ge, B. (2023). Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852*.
McFadden, T. U., & Gillam, R. B. (1996). An examination of the quality of narratives produced by children with language disorders. *Language, Speech, and Hearing Services in Schools*, 27(1), 48-56.
McGilvray, J. (2014). *Chomsky: Language, mind and politics.* Polity.
Mikk, J. (2008). Sentence length for revealing the cognitive load reversal effect in text comprehension. *Educational Studies*, 34(2), 119-127.
Nasser, A. N. A. (2016). Teaching the writing skill to Yemeni EFL learners: The importance and challenge. *South-Asian Journal of Multidisciplinary Studies (SAJMS)*, 3(6), 191-203.
Nippold, M. A. (2000). Language development during the adolescent years: Aspects of pragmatics, syntax, and semantics. *Topics in language disorders*, 20(2), 15-28.
Oatley, K. (1995). A taxonomy of the emotions of literary response and a theory of identification in fictional narrative. *Poetics*, 23(1-2), 53-74.
Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., ... & Tao, D. (2023). Towards making the most of chatgpt for machine translation. *arXiv preprint arXiv:2303.13780*.
Piaget, J. (1976). Piaget's theory.
Pinker, S. (2003). *The language instinct: How the mind creates language.* Penguin UK.
Prado, J., Spotorno, N., Koun, E., Hewitt, E., Van der Henst, J. B., Sperber, D., & Noveck, I. A. (2015). Neural interaction between logical reasoning and pragmatic processing in narrative discourse. *Journal of cognitive neuroscience*, 27(4), 692-704.
Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*.
Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?. *Journal of Applied Learning and Teaching*, 6(1).
Shahriar, S., & Hayawi, K. (2023). Let's have a chat! A Conversation with ChatGPT: Technology, Applications, and Limitations. *arXiv preprint arXiv:2302.13817*.
Shidiq, M. (2023, May). The use of artificial intelligence-based chat-gpt and its challenges for the world of education; from the viewpoint of the development of creative writing skills. In *Proceeding of International Conference on Education, Society and Humanity* (Vol. 1, No. 1, pp. 353-357).

Sun, M. (2023). The Language Errors and Their Correction Methods in Middle-Grade Elementary School Composition. *A Successful Way to Composition*, (15), 38-40.

Tan, Y., Min, D., Li, Y., Li, W., Hu, N., Chen, Y., & Qi, G. (2023). Evaluation of ChatGPT as a question answering system for answering complex questions. *arXiv preprint arXiv:2303.07992*.

Tang, H. (2018). Training in the correct use of punctuation marks for primary school students. *Culture Study*, (7).

Tong, X., Mo, J., Shu, H., Zhang, Y., Chan, S., & McBride-Chang, C. (2014). Understanding Chinese children's complex writing: Global ratings and lower-level mechanical errors. *Writing Systems Research*, 6(2), 215-229.

Transformer, C. G. P. T., & Zhavoronkov, A. (2022). Rapamycin in the context of Pascal's Wager: generative pre-trained transformer perspective. *Oncoscience*, 9, 82.

Vaishya, R., Misra, A., & Vaish, A. (2023). ChatGPT: Is this version good for healthcare and research?. Diabetes & Metabolic Syndrome: *Clinical Research & Reviews*, 17(4), 102744.

Wagner, R. K., Puranik, C. S., Foorman, B., Foster, E., Gehron, L., Tschinkel, E., & Kantor, P. T. (2011). Modeling the development of written language. Reading and Writing: *An Interdisciplinary Journal*, 24(2), 203–220.

Wang, R. (2021). Issues and Solutions in Teaching Introductory Composition for Third-Grade Language Arts. *Teacher's Journal* (S1), 156.

Wang, Y. (2023). An Analysis of the Causes of Typographical Errors in Third-Grade Elementary Students and a Study on Teaching Strategies (Master's thesis, Chongqing Three Gorges University).

Xu, D. (2018). Strategies for Teaching Narrative Writing to Third Graders in Elementary School. *The test and study*, (35), 126.

Yan, C. M. W., McBride-Chang, C., Wagner, R. K., Zhang, J., Wong, A. M. Y., & Shu, H. (2012). Writing quality in Chinese children: Speed and fluency matter. *Reading and Writing: An Interdisciplinary Journal*, 25, 1499–1521.

Yuan, Y. (2023). Theoretical Reflections on Linguistic Studies Against the Background of AI Great Leap Forward. *Chinese Journal of Language Policy and Planning,* 8(4), 7-18.

Zhang, Z. (2023). A Study on the Methods of Teaching Composition for Third-Grade Primary School under the New Curriculum Reform. *GUOJIA TONGYONG YUYANWENZI JIAOXUE YU YANJIU*, (02), 167-169.

Zhao, W., Strube, M., & Eger, S. (2022). DiscoScore: Evaluating text generation with. BERT and discourse coherence. arXiv preprint arXiv:2201.11176.

Zhao, W., Zhao, Y., Lu, X., Wang, S., Tong, Y., & Qin, B. (2023). Is ChatGPT Equipped with Emotional Dialogue Capabilities?. a*rXiv preprint arXiv:2304.09582.*

Zhou, T., Cao, S., Zhou, S.*, & He, A*.(2023). Chinese intermediate English learners outdid ChatGPT in deep cohesion: Evidence from English narrative writing. *System*, 118.